University of Alberta

Development of a Spectral Searching Strategy for Peptide and Protein Identification

by

Mingguo Xu

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

Fall 2012
Edmonton, Alberta

Abstract

The overall goal of this thesis research is to develop a spectral searching strategy capable of identifying peptide sequences from MS/MS spectra with high sensitivity and accuracy.

First, a shotgun proteome analysis method was developed and successfully applied to the identification of proteins from thousands of cancer cells. This work illustrated that proteome profiling of a small number of cells isolated from blood can be achieved. By comparing the obtained profile to a standard profile, cell typing might also be possible. This method may prove to be useful for cancer diagnosis or prognosis. From this study, we realized that sequence database searching strategy is one of the bottlenecks to achieve better sensitivity of protein identification for proteome profiling work.

As a promising alternative, spectral searching strategy is believed to be able to provide more sensitive and accurate peptide and protein identification. In spectral searching strategy, there are two main components: spectral libraries and the searching algorithm.

Since an accurate identification by spectral searching strategy is built on the premise of a reliable MS/MS spectral library, ^{15}N-metabolic labeling and ^{18}O-labeling approaches were developed to experimentally validate all the peptide matches from sequence database search results.

With those validated matches, the sensitivity and accuracy of commonly used search engines (Mascot and X!Tandem) and two popular statistical approaches (PeptideProphet and Percolator) were carefully examined. Moreover, two strategies were designed to identify single-hit protein identifications (proteins identified by only one peptide) with high reliability. In addition, Percolator was successfully interfaced with X!Tandem to enhance its performance.

Finally, a spectral searching algorithm called SpecMatching was developed to utilize the experimentally validated spectral library. In analyzing a digest of an *E. coli* extract using both Mascot and SpecMatching, it was shown that SpecMatching provided better sensitivity and specificity even with this small-size spectral library.

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor, Dr. Liang Li, for granting me the opportunity to study in his research group and for his invaluable advice, supervision, guidance and encouragement throughout my research. I have learnt a great amount from him and I am sure it will continue to benefit my life and career.

I would like to gratefully acknowledge the other members of my supervisory committee, Dr. Robert E. Campbell, Dr. Derrick L.J. Clive and the other members of my thesis examining committee, Dr. Guohui Lin, Dr. Michael J. Serpe and Dr. David M. Lubman for their active participation during my oral examination, their thorough reviews and comments on this thesis, and their valuable advice on my research.

I would like to thank all the members of Dr. Li's research group, especially Dr. Avalyn Stanislaus for her continuous advice and great support, as well as for being so encouraging and helpful. I would also like to thank Dr. Azeret Zuniga for her invaluable wisdom and humorous personality, making this whole experience more enjoyable. My appreciation also goes to Dr. Nan Wang and Dr. Andy Lo for their professional training and wise words about tackling a PhD program one day at a time. Many thanks to Zhendong Li for his help on data analysis and program development.

Special thanks go to Dr. Randy Whittal, Béda Reiz and Jing Zheng from

the Mass Spectrometry Facility in the Department of Chemistry at the University of Alberta for sharing some of their expert mass spectrometry knowledge and providing excellent technical assistance in analyzing samples on the Q-TOF. I must also thank Dr. Fang Wu, Dr. Sandra Marcus and Dr. Garreth Lambkin for their contribution in cell culture, as well as Dr. Peng Wang for his great help in flow cytometric analysis.

I would also like to thank the departmental staff in the general and purchasing offices, the mailroom, and the electronics and machine shops for their kind help.

I would like to acknowledge Fs Chia PhD Scholarship, the Department of Chemistry and the University of Alberta for providing me with financial support while completing my graduate studies.

Many thanks also go to my friends for their friendship, love and encouragement in the pursuit of my degree.

Finally, I especially thank my parents, Qingming Xu and Guo Li, for their selfless love, patience and encouragement over the years. It is their love and confidence in me that have shaped me into the person I am today. I will be forever grateful.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D	Two-dimensional
^{o}C	Degree Celsius
%	Percent
%B	Organic solvent composition
AC	Alternating current
ACN	Acetonitrile
ALS	Acid labile surfactant
AnnoPeaks	The fraction of high intensity peaks being annotated as fragment ions
BSA	Bovine serum albumin
CE	Collision energy
CID	Collision-induced dissociation
CHCA	α-cyanohydroxycinnamic acid
CL	Confidence level
Da	Dalton
DC	Direct current
DDA	Data directed analysis
DeltaM	The difference in calculated and observed mass (Th)
DeltaScore	The difference of HyperScore between the best and the second best peptide matches
DHB	2, 5-dihydroxybenzoic acid

DNA	Deoxyribonucleic acid
DTT	Dithiolthreitol
E-value	Expectation value
EDTA	Ethylenediaminetetraacetic acid
EnzC	Boolean value: does the peptide have a C-terminal enzymatic (tryptic) site?
EnzN	Boolean value: is the peptide preceded by an enzymatic (tryptic) site?
EPA	Environmental Protection Agency
ESI	Electrospray ionization
FA	Formic acid
FDR	False-discovery rate
FragError	The average mass error of all the fragment ions
FSC	Forward scatter
FT-ICR	Fourier-transform ion cyclotron resonance
g	Relative centrifugal force
GC	Gas chromatography
GC-MS	Gas chromatography-mass spectrometry
GPMDB	Global Proteome Machine Database
h	Hour
HEA	Human epithelium antigen
HCl	Hydrochloric acid
HPLC	High performance liquid chromatography

HSH	Homologous single-hit
HyperScore	X!Tandem's score for the peptide Identification
IAA	Iodoacetamide
IonFrac	The fraction of fragment ions being matched in an ion series
IonNo	The number of peaks that matched between the theoretical and the test mass spectrum
IonScore	The summed intensities of different types of fragment ions
IPI	International protein index
KR PSM	Peptide matches with C-terminal lysine or arginine
L	Litre
LB	Lysogeny broth
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
LCM	Laser capture microdissection
Log(E)	The $\log_{10}$ value of the expectation value for the peptide match
m	Mass
m/z	Mass-to-charge
M	Molar
MALDI	Matrix-assisted laser desorption/ionization
MaxI	Maximum fragment ion intensity
min	Minute
MissClea	The number of missed internal enzymatic (tryptic) sites
ModFrac	The number of variable modifications

ModNo	The number of variable modifications
MRM	multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
Multi-hit	Protein identification with at least two unique peptides
n	The number of candidate peptides
NextScore	The HyperScore of the second best peptide match of the spectrum
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NP40	Nonidet-P40
p	Probability
PBL	Peripheral blood leukocytes
PBS	Phosphate-buffered saline
PEG	Polyethylene glycol
PEP	Posterior error probability
PepLeng	The length of the peptide identification
pH	Potential of hydrogen
PIE	Precursor ion exclusion
PMF	Peptide mass fingerprinting
ppm	Parts per million
PSM	Peptide-spectrum match
PSMSumI	The $\log_{10}$ value of the sum of all of the fragment ion intensities

Q-TOF	Quadrupole time-of-flight mass spectrometer
R^2	Square of the correlation coefficient
RelDeltaM	The relative difference in calculated and observed mass (ppm)
RF	Radiofrequency
ROC	Receiver-operating characteristic
RP	Reversed phase
RPLC	Reversed phase liquid chromatography
s	Seconds
S/N	Signal-to-noise ratio
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
Single-hit	Protein identification with only one unique peptide
SRM	Selected reaction monitoring
SSC	Side scatter
SSH	Strict single-hit
SVM	Support vector machine
t	Time
TCA	Trichloroacetic acid
TFA	Trifluoroacetic acid
TFE	Trifluoroethanol
Th	Thomson
TOF	Time-of-flight

Tris	Tris (hydroxymethyl) aminomethane
U	DC voltage
UV	Ultraviolet
V	Voltage
XT	X !Tandem
XP	X !Tandem Percolator

Chapter 1

Introduction

1.1 Proteomics

Originating from the words protein and genome, the term "proteome" represents the total collection of proteins encoded by the genes in one organism.[1, 2] Proteomics, therefore, is defined as the study of the entire protein content, such as the understanding of the structure, interaction, and function of all the proteins within an organism.[3] One of the major causes of diseases is defective proteins, in turn making proteins as useful indicators for the diagnosis of a particular disease. In addition, proteins are the primary targets of most drugs and thus are the main basis for the development of new drugs. Therefore, the study of proteome is important for understanding their role in the cause and control of diseases. At present, the aims of proteomics may be roughly categorized into four directions: (1) large-scale protein identification and their post-translational modifications (PTMs); (2) differential expression analysis of proteins in healthy and diseased states; (3) studies of protein interactions; and (4) studies of protein functions.[4] In any of these scenarios, knowing the protein sequence is fundamental to understand the roles of proteins in biological processes on the molecular level.

1.2 Mass spectrometry Based Protein Sequencing

Due to recent advances in instrumental analysis and bioinformatics, mass spectrometry has become the leading technique to sequence proteins. It mainly involves four stages: protein sample preparation, ionization, mass spectrometric analysis and data interpretation.

1.2.1 Sample Preparation

Most samples for proteomics experiments nowadays are derived from cellular samples (e.g., cultured cells and primary cells). In this case, protein extraction is often required. It involves disrupting the cellular membrane using a combination of physical actions (e.g., sonication or pressure) and extraction buffers capable of rupturing cellular structure (e.g., NP-40). Removal of lipids, surfactants and other small molecules is needed to avoid interference with downstream mass spectrometric analysis of proteins. This goal can be effectively achieved with solvent[5, 6] or trichloroacetic acid (TCA) protein precipitation[7]. For instance, when acetone is added to an aqueous protein solution, the solubilized proteins are denatured and form precipitates.

Once reasonably pure protein precipitates are obtained, complete solubilization of proteins is often required. It involves the use of a solvent or a series of solvents to disrupt the protein-protein interactions caused by van der Waals forces, electrostatic forces and hydrogen bonding. Many efforts have been devoted to the development of effective and efficient protein solubilization.[8] A

variety of reagents, such as aqueous solution (e.g., ammonium bicarbonate), organic solvent (e.g., methanol), chaotropic agents (e.g., urea) and surfactants (e.g., SDS), have all been carefully studied for their protein solubilization capability.[9-11]

Next, in order to reduce the complexity of protein mixture, protein separation is often performed. As the most widely used protein separation approach, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) separates proteins based on their molecular weight or size. SDS-PAGE is noted for its high resolution and fair loading capacity.[12] However, the subsequent sample workup steps, such as protein extraction, purification and digestion within gel pieces, can be laborious and inefficient. Apart from the gel-based separation approach, solution based separation methods, such as protein reversed phase liquid chromatography (RPLC)[13] and affinity chromatography[14], operate in different manners. They offer the advantage that proteins stay in solution, thus making the subsequent sample workup procedures on both protein and peptide level less tedious.[15]

In bottom-up proteomics[16] (analyzing peptides that result from protein digestion), protein samples are digested into smaller peptides using enzymes[15] (e.g., trypsin, chymotrypsin or pepsin) or chemical methods (e.g., cyanogen bromide[17] and microwave-assisted acid hydrolysis[18]) before mass spectrometric analysis. Of all the digestion methods, the most common one utilizes trypsin. Trypsin cleaves proteins at the carboxyl side of lysine or arginine, except when

either is followed by proline. Its popularity can be attributed to three reasons. First, the high specificity of trypsin provides reproducible digestion results. Secondly, lysines and arginines are distributed along the protein sequence in spaces that, trypsin digestion results in peptides (tryptic peptides) that have molecular weights ranging from 600 to 3000 Da, which is ideal for mass spectrometric analysis. Finally, most tryptic peptides contain a basic amino acid, either lysine or arginine, at the C-terminus. It results in tryptic peptides that are readily protonated in electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) process (see section 1.2.2). Ionization efficiency of tryptic peptides is generally higher than peptides containing no lysine or arginine.

After digestion, the complex mixture of peptides can be further fractionated using various separation methods. Due to the compatibility of most mobile phase solvents and additives to electrospray ionization, reversed phase liquid chromatography (RPLC) is the most common and typically last dimension of separation that is coupled with a mass spectrometer.[15, 19] In RPLC, the stationary phase surface contains a non-polar alkyl chain with a silane linkage to the stationary phase such as C18. Peptides can be retained on the stationary phase via hydrophobic forces. The elution of peptides is achieved by decreasing the polarity of solvent mixtures in the mobile phase (e.g., reducing the water content while increasing acetonitrile). Meanwhile, strong cation exchange (SCX) at low pH (i.e., pH < 3.0) is also commonly used for fractionation of tryptic peptides[19], since the N-terminus amine or basic side chain of lysine or arginine ensures retention of most tryptic peptides. In SCX separation, the functional group of

stationary phase is anionic, such as the $-SO_3^-$ group. At acidic conditions (i.e., pH < 3.0), most peptides in the solution will be positively charged and thus interact with the SCX stationary phase via ionic interaction. Peptide elution is performed by increasing the salt concentration in the mobile phase (e.g., increasing the concentration of KCl in the elution solution).

In Chapter 2, RPLC on the peptide level was used to separate peptide mixtures before mass spectrometric analysis. In Chapter 3, an off-line SCX-RPLC configuration was adopted to simplify peptide mixtures. In Chapter 4, protein level RPLC fractionation was first applied to protein mixtures from human cells lysates. After digestion, RPLC-MS/MS was performed on the tryptic peptides.

1.2.2 Ionization Methods

In order to perform mass spectrometric analysis on peptides or proteins, they must first be ionized. Since peptides or proteins are thermally unstable, a soft ionization technique is required. There are two major approaches for ionization of proteins and peptides, electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI).

Electrospray ionization[20] begins when the LC eluent is sprayed through a conducting capillary under the influence of a high voltage, typically between 2 to 5 kV. The schematic diagram of ESI process is described in Figure 1.1. Because of the high electric field at the capillary tip, cations concentrate at the capillary tip and anions migrate away from the tip. Thus, cations get enriched at the surface of

the liquid meniscus, forming a Taylor cone. Provided with a sufficiently high voltage, the tip of the cone becomes unstable and breaks into a fine jet, which contains numerous small charged droplets. The solvent of the charged droplets then evaporates, often aided by a heated source region and/or a flow of dry gas (e.g., N_2). As the solvent evaporates, the excessive surface charges begin to repel the charged analytes on the droplet surface. Once the charge density exceeds the Rayleigh limit, the charged analytes reside on the surface of the droplet that then evaporate and escape into the atmosphere and become gas phase ions. Consequently, it stabilizes the charged droplet. As the solvent evaporates further, this process is repeated to produce more gas phase ions. This proposed mechanism is called ion evaporation model[21].

Alternatively, the charged residue model[22, 23] predicts the formation of smaller fission droplets from the main droplet. Solvent evaporation of the main droplet leads to an increase in charge density. When the repulsive forces of the charges exceed the droplet surface tension, Coulomb fission happens, where the main droplet produces small charged progeny droplets. If there is an analyte molecule within this smaller droplet, continuing desolvation eventually leads to charge transfer to the analyte and formation of the gas phase ion.

Depending on the type of analytes, it has been suggested that either model may be better at rationalizing the formation of gas phase ions. Since electrospray ionization is a competitive process, the chemical properties of the analyte (e.g., hydrophobicity and gas phase basicity) directly affect its chance of ionization.

The ion evaporation model well explains why having basic residue(s) in a peptide sequence is beneficial for mass spectrometric analysis.

In MALDI[24], the ionization of peptides/proteins has three steps. First, analytes are acidified to provide protonation. Second, a matrix is mixed with analytes at certain ratios (e.g., 1000 : 1). The commonly used matrices in MALDI are α-cyanohydroxycinnamic acid (CHCA) and 2, 5-dihydroxybenzoic acid (DHB). CHCA is a pale yellow solid that forms a uniformly flat sample spot, whereas DHB is white solid that forms small needle-like crystals on the target. Depending on the analytes, one matrix may give better signal response or cleaner spectra than the other. Since the ionization mechanism in MALDI has not been perfectly understood, the choice of matrix is generally determined empirically. Then upon shining of an UV laser beam at the sample spot, analyte molecules are lifted from the MALDI plate by matrix material into the gas phase and ionized. As an alternative ionization method for peptides and proteins, MALDI has some unique merits, such as higher tolerance to salt concentration and surfactants. However, the difficulty of interfacing with separation techniques (e.g., LC) makes it less popular than ESI in proteomic studies.

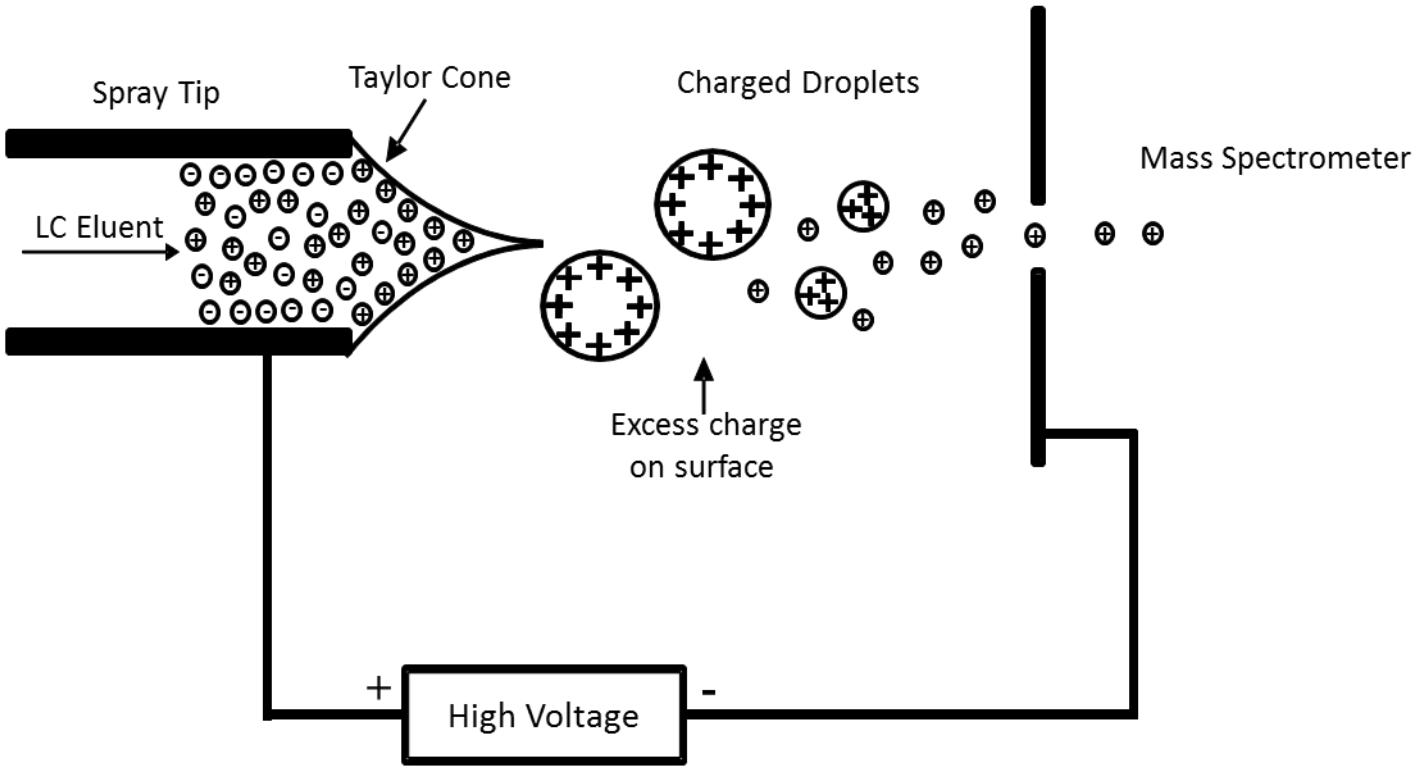

Figure 1.1 Schematic of the electrospray ionization process.

1.2.3 MS Instrumentation

After ionization, analyte ions are transported into a mass analyzer for MS and/or MS/MS analysis. There are a variety of mass analyzers available that can be used for proteomic studies, such as time-of-flight (TOF)[25], Fourier-transform ion cyclotron resonance (FT-ICR)[26] and Orbitrap[27]. The main instrument used for this thesis work was quadrupole time-of-flight mass spectrometer (Q-TOF), a hybrid mass spectrometer combined from quadrupole and TOF mass analyzers. Its instrumentation will be discussed in detail.

The quadrupole[28] mass analyzer is constructed from four parallel cylindrical metal rods (see Figure 1.2). Two diametrically opposed rods are paired up. A potential of $(U + V \cos(\omega t))$ is applied to one pair and a potential of $-(U + V \cos(\omega t))$ is applied to the other pair. The quadruple mass analyzer can play two different roles, an ion guide or mass filter, depending on the specific application of AC/DC on the rods. In the RF-only mode, the DC component is set to zero ($U = 0$) and ions of a large range of m/z values can be successfully transmitted through the quadrupole. In this mode, the quadrupole acts as an effective ion guide. When the DC component is not set to zero ($U \neq 0$), by applying a specific DC and AC voltage, only ions of a certain *m/z* value have a stable trajectory. All the other ions are lost in the transition. In this mode, the quadruple acts as a narrow band mass filter.

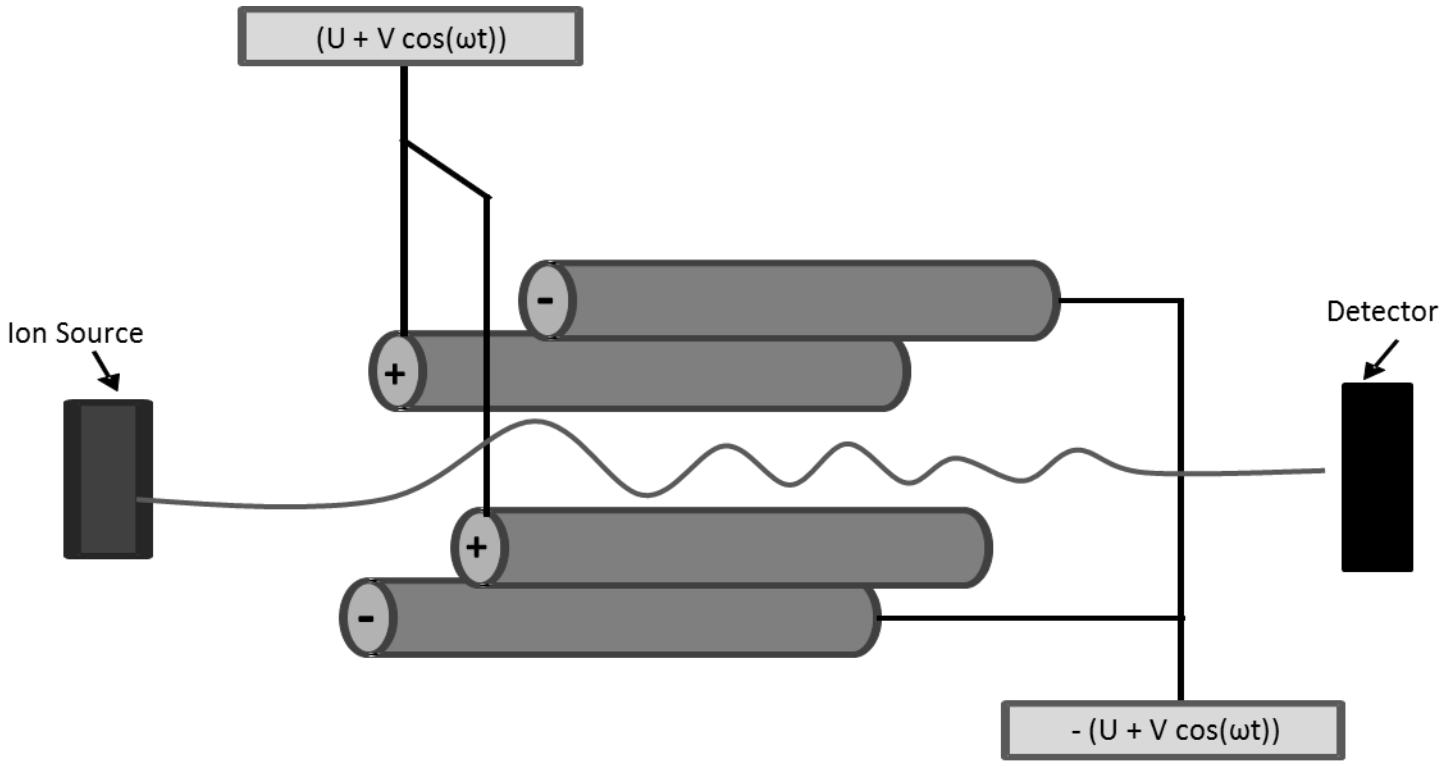

Figure 1.2 Schematic diagram of a quadrupole mass analyzer.

Because quadrupoles can be used as both ion guides and narrow band filters, they are commonly seen as a key component in hybrid mass spectrometers. Figure 1.3 displays the schematic of the Q-TOF premier systems from Waters. This system consists of an ESI source, a quadrupole unit, a collision cell and a TOF mass analyzer.

In the TOF mass analyzer, ions are pulsed in the direction of the flight path with minimum distance dispersion by an extraction voltage. The velocity of an ion (v) can be defined as:

$$v = \sqrt{\frac{2eV}{m}}$$

where *V* is the voltage, *e* is the charge of the ion, *m* is the mass of the ion. The velocity of an ion is inversely proportional to its *m/z* value. Therefore, lower *m/z* ions reach the detector earlier when traveling the same distance. The time an ion takes to reach the detector can be calculated by:

$$t = L\sqrt{\frac{m}{2eV}}$$

where *L* represents the length of the ion path. Thus, the measured arrival times of ions can be readily converted to *m/z* values, consequently constructing a mass spectrum.

In a modern TOF mass spectrometer, a reflectron[25] is used to compensate the initial spatial and kinetic energy dispersion of analyte ions to achieve better

resolution. The reflectron is usually placed at the end of the flight tube and consists of a series of grids and electrodes within which an electric field gradient is created. As ions with different kinetic energy enter the field, the ones with higher energy will penetrate deeper into the reflectron, increasing their flight path and observed flight time. As a result, ions with the same *m/z* value but slightly different initial kinetic energies will eventually arrive at the detector simultaneously. The net effect of the implementation of a reflectron is improved mass resolution ($m/\Delta m = \sim 10{,}000$).

The Q-TOF Premier system combines the merits from both the quadrupole and TOF mass analyzers, allowing automated accurate mass measurement of both precursor and fragment ions. Benefitting from the high ion transmission efficiency of ZSpray™ source technology and real time mass calibration capability of NanoLockSpray™ technology, the Q-TOF Premier system offers sensitive and accurate mass measurement. Mass accuracy of 30 ppm or less is routinely achieved. The Q-TOF premier system operates in three different modes: MS survey mode, MS/MS mode and data directed analysis (DDA) mode. In MS survey mode, the quadrupole unit works as an ion guide, allowing a wide range of ions to be transmitted. The *m/z* values and intensities of those ions are measured in the TOF analyzer to generate a mass spectrum. In MS/MS mode, resolving DC is applied on the quadrupole. The quadrupole unit operates as a narrow mass filter to isolate candidate ions for fragmentation. The selected ions are fragmented in the collision cell and all the fragment ions are transported to the TOF mass analyzer. An MS/MS spectrum is then recorded. In the DDA mode, the instrument

is set to automatically switch between MS and MS/MS modes depending on the ions detected during the MS survey mode.

In this thesis work, the DDA mode was chosen for most of mass spectrometric analysis on peptides. It enables intelligent MS and MS/MS analyses to be performed automatically, maximizing the amount of real information acquired on components of interest. By using this mode, the mass spectrometric analysis can be targeted for analytes with specific charge state, *m/z* value etc., making the development precursor ion exclusion[29] (PIE) and inclusion strategies possible (see Chapter 4).

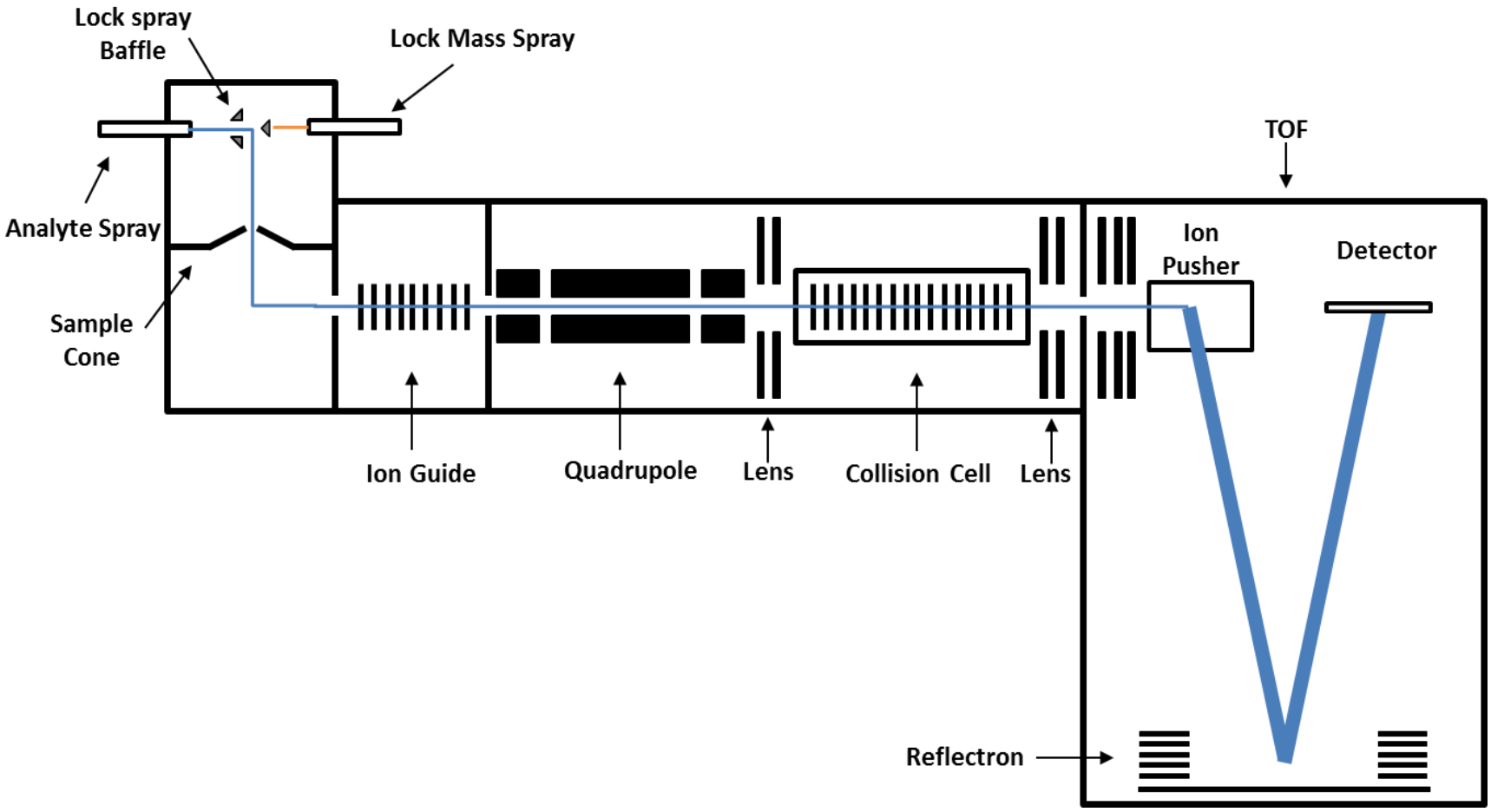

Figure 1.3 Schematic of Waters Q-TOF premier system.

1.2.4 Tandem Mass Spectrometry

In the DDA mode of Waters Q-TOF premier system, a duty cycle during mass spectrometric analysis of peptides is initiated by the acquisition of an MS spectrum (MS survey scan). Signals are then quickly processed and the most intense peaks are selected for MS/MS fragmentation. Then the system is set to MS/MS mode to collect an MS/MS spectrum. The MS/MS spectral acquisition begins when precursor ions selected by the quadrupole enter the collision cell to be fragmented. The collision cell is a hexapole or octopole within a set of acceleration plates. In order to accelerate the ions through the cell, a slight voltage is applied across the plates. In the cell, collisions with a neutral, inert bath gas (e.g., nitrogen or argon) increases the internal energy within the ions, converting part of the kinetic energy into internal energy of the ions, thus resulting in bond fragmentation. This process is referred as collision-induced dissociation (CID). These resultant fragment ions are then measured by the TOF mass analyzer to record an MS/MS spectrum.

In practice, CID can be performed with either high or low collision energy. Low energy CID (10-100 eV) is widely used in most mass spectrometric proteome analysis. During low energy CID, the fragment ions of a peptide precursor ion are mainly produced by the breakdown of the peptide backbone.[30] Along the peptide backbone there are several bonds that can be broken during fragmentation, rendering different types of fragment ions (see Figure 1.4). The most commonly observed fragment ions are b- (see Figure 1.5A) and y- ions (see

Figure 1.5B), which are formed from the cleavage of the C-N bond in the amide backbone with charge retention on the N- or C-terminus, respectively. Occasionally, a-ions (see Figure 1.5C) and immonium ions (see Figure 1.5D) are also observed in the low energy CID process. Neutral losses of water (-18.011 Da) can be observed for fragment ions containing threonine, serine, glutamic acid or aspartic acid. For fragment ions containing lysine, arginine, glutamine or asparagine, the neutral losses of ammonia (-17.027 Da) can sometimes be observed.

This thesis work focused on the mass spectrometric analysis of tryptic peptides. They display a favorable fragmentation pattern in the CID process due to the present of a C-terminal basic amino acid residue (i.e., lysine or arginine). Dominant y- fragment ion series is commonly witnessed for most tryptic peptides, in comparison with the observation that non-tryptic peptides with random locations of basic amino acid residues give rise to a mixture of b- and y- fragments. By placing the basic residues at the C-terminus, peptides fragment in a favorable and more predictable manner throughout the entire sequence, which facilitates the elucidation of the sequence information.[31]

With all the fragmentation rules available, one can start sequencing peptides or proteins based on the fragment ions recorded in the MS/MS spectra. This task can be done either manually or by software. In the next section, the interpretation of mass spectrometric data will be discussed in detail.

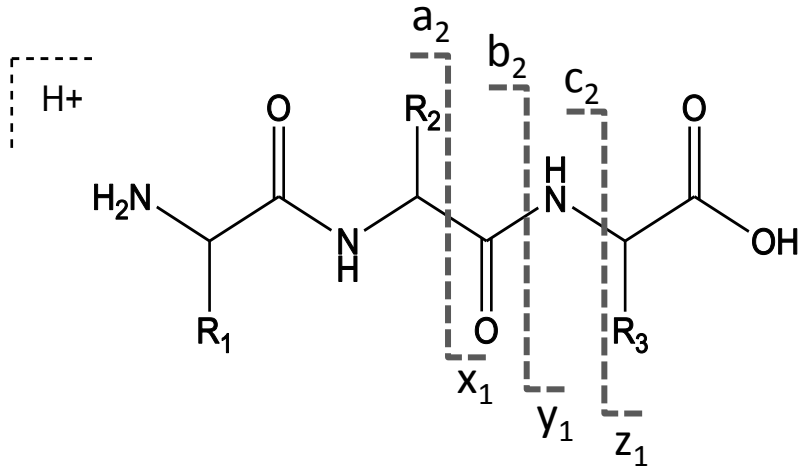

Figure 1.4 Product ion nomenclature.

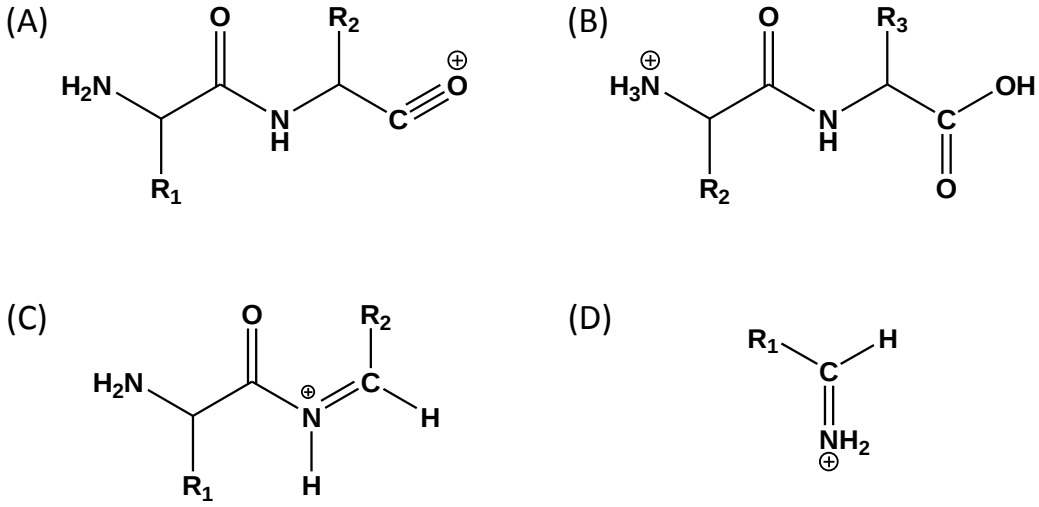

Figure 1.5 An example of (A) b ions, (B) y ions, (C) a ions and (D) immonium ions.

1.3 Data Interpretation

The two major mass spectrometric strategies of protein sequencing are (1) MS analysis of relatively short peptides obtained from the analyte proteins by enzymatic or chemical reactions (e.g., trypsin digestion or acid hydrolysis) and (2) sequencing by MS/MS analysis of selected precursor ions (peptides or proteins) and predictable fragmentation patterns associated with amino acid sequences.

1.3.1 Peptide Mass Fingerprinting

A peptide mass fingerprint is the collective mass measurements of the peptides derived from a protein upon defined enzymatic or chemical cleavages (e.g., trypsin digestion). The concept of peptide mass fingerprinting (PMF) as a rapid and reliable approach for protein identification is based on the fact that the set of masses for peptides produced by residue specific enzymatic or chemical digestion is unique to any given protein.[32] A typical workflow of PMF experiment is illustrated in Figure 1.6. In PMF, protein identification is accomplished by matching the observed peptide masses to the theoretical masses derived from the proteome database using a search engine. There are a variety of search engines available such as Mascot (http://www.matrix-science.com), MS-FIT (http://prospector.ucsf.edu/prospector/mshome.htm) and PeptideMapper (http://www.nwsr.manchester.ac.uk/mapper/). However, the presence of numerous proteins in a mixture can significantly complicate the identification process, rendering unreliable results. Therefore, in PMF, protein separation techniques (e.g., SDS-PAGE) are often required to simplify protein contents.[33, 34]

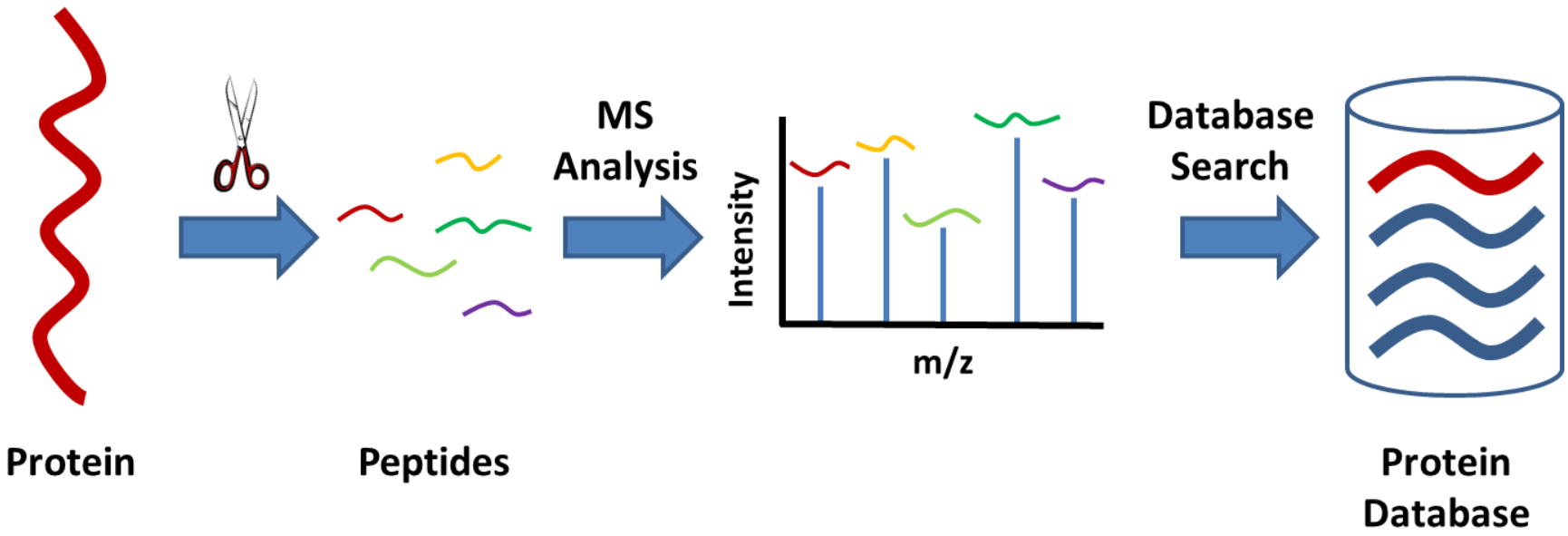

Figure 1.6 Peptide mass fingerprinting workflow.

1.3.2 Sequence Database Search

Unlike MS analysis, tandem mass spectrometric analysis of peptides can provide the information of molecular weights of peptides, as well as the amino acid arrangement within. It is all attributable to the predictable fragmentation pathways of peptides, which usually results in the production of characteristic amino- and carboxyl-terminus containing fragments (see Figure 1.4). Moreover, another advantage of the MS/MS analysis is that sequence information is often obtainable from highly complicated peptide mixtures because the individual precursor ions are isolated in the first stage of the tandem operation. Advanced instrumentation of modern mass spectrometers allows for acquisition of both MS and MS/MS data in the same analysis through data-directed switching (see section 1.2.3). Thanks to these advantages, strategies based on MS/MS analysis, such as *de novo* sequencing[35], sequence tag searching[36] and sequence database searching[37-39], have attracted much more attention. Of all those strategies, sequence database search is the method of choice for most proteomic studies due to its relatively fast speed and good accuracy. Sequence database search is used in this thesis work.

In a typical sequence database search, *m/z* value of the precursor and its charge state are used to calculate the experimental mass of the peptide. Within a defined mass error tolerance, search engines find all possible peptide candidates with a similar peptide mass from a sequence database. Only those peptide candidates will be further compared with experimental data. Next, theoretical

fragmentation patterns are generated for each of these peptide candidates. By comparing the theoretical fragmentation pattern with the experimental MS/MS spectrum, peaks in the experimental spectrum can be annotated with different fragment ions. Each of the potential matches is then valued according to how well the theoretical fragmentation pattern accounts for peaks in the experimental MS/MS spectrum (see Figure 1.7).

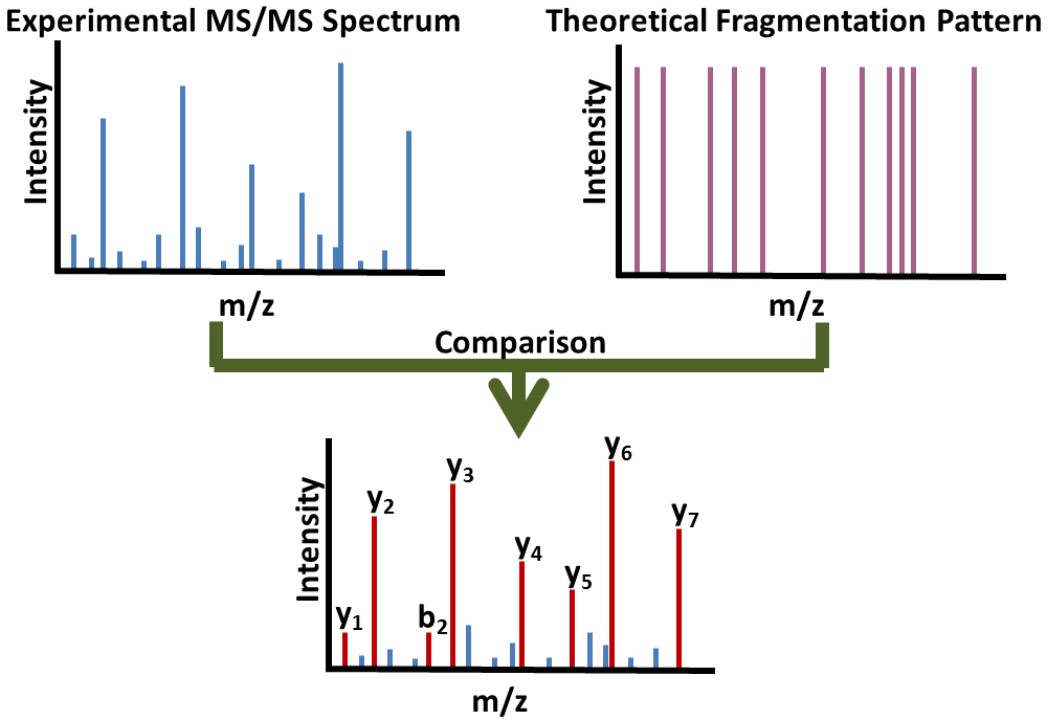

Figure 1.7 Peptide sequencing by annotating MS/MS spectra through database searching.

Once the peptide sequences are identified, protein identifications can be inferred from peptide match information. Theoretically, a protein is considered identified when at least one of its associated peptides is found (assuming the same peptide sequence is not present elsewhere in the proteome). When complex protein mixtures are analyzed all together, the linkage between peptides and

proteins is lost, which makes the protein inference task challenging. Nonetheless, most search engines, such as Mascot and X!Tandem, are capable of intelligently grouping peptides to render protein identifications. Even though they differ in practice, all the algorithms aim at deriving the simplest list of proteins sufficient to explain the observed peptides (Occam's razor).

1.3.2.1 Mascot

Mascot[37] (Matrix Science, London, UK) is a commercially available software that uses mass spectrometry data to identify proteins from primary sequence databases. In Mascot, a probability-based scoring scheme is adopted to evaluate how reliable peptide-spectrum matches (PSMs) are. It assumes that a match between the experimental data and each sequence entered in the database is a chance event. Therefore, the match with the lowest probability (p) of being a random occurrence is considered as the best fit. Then Mascot ion score is calculated as $-10\times\log_{10}(p)$. Moreover, in order to compensate the problem of multiple comparisons, Mascot identity threshold is implemented. In the definition of Mascot identity threshold, $-10\times\log_{10}(p/n)$, p is the defined error rate and n represents the number of candidate matches. The match with the Mascot ion score above the threshold is generally considered to be a significant peptide assignment. For example, if there are 500 candidate PSMs and one is comfortable with a 1 in 20 chance of getting a false positive match (an error rate of 0.05), the Mascot identity threshold would be 40. A PSM with Mascot ion score above 40 will be considered a significant peptide assignment. Even though the detailed algorithm

of Mascot is not released, it is still considered one of most powerful search engines in the proteomic field.[40]

1.3.2.2 X!Tandem

X!Tandem[38] is another popular search engine capable of matching MS/MS data with peptide sequences. Unlike Mascot, X!Tandem is an open source program. Instead of reporting a probability and identity threshold for each PSM, X!Tandem adopted the concept of reporting expectation values (E-values) of PSMs. In its algorithm, X!Tandem first measures the spectral similarity between the experimental spectrum and several candidate theoretical peptide fragmentation patterns, generates hyperscores (the sum of matched fragment ion intensities multiplied by the N factorial for the number of matched ions), plots a distribution of hyperscores for the spectral search and extrapolates an E-value to provide a statistical evaluation for each identification. E-value is defined as the number of random matches that would be expected to have the same or better scores. This X!Tandem scoring scheme is an empirical measure of whether the match is an outlier (see Figure 1.8). Mascot also implemented this idea in their later version of the software, called Mascot homology threshold. It is believed that this empirical scoring scheme might offer better sensitivity especially when searching a large space.[41] However, due to the different ways of performing statistical analysis, expectation values (E-values) from X!Tandem are not directly comparable with the Mascot ion score.

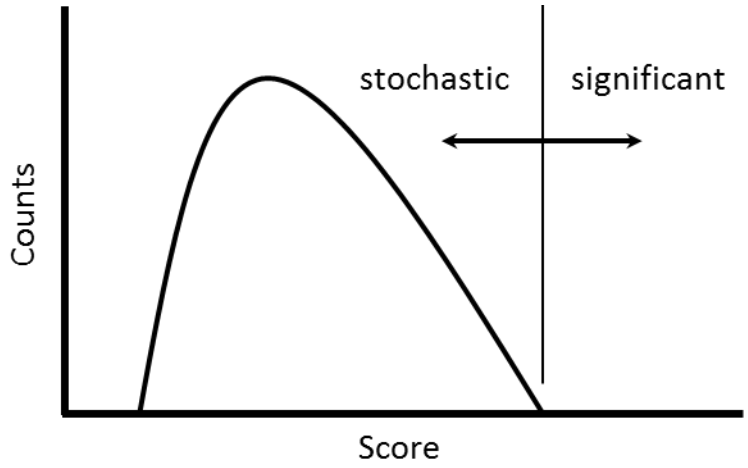

Figure 1.8 Schematic representation of a stochastic score distribution. Any PSM with a score inside the body of the stochastic distribution is not seen as valid identification. A score higher than the right-hand boundary of the stochastic distribution may be assigned as potential valid, with an associated expectation value. Adapted from Fenyo and Beavis. [38]

1.3.2.3 Statistical Analysis

Scores, either probabilities or expectation values, are always needed to evaluate the validity of peptide sequence matches in any search engine result. While it provides statistical assessment of individual PSM, it fails to compute error estimates for the collection of peptide identifications. During the past decade or so, the concept of global false-discovery rates (FDRs) has emerged, matured and been widely accepted in the field of proteomics.[42-44] It is defined as the fraction of false positives in all the positive identifications. However, how to accurately estimate the global FDR for a search result remains to be a challenging problem. Numerous strategies, such as the target-decoy approach[45], PeptideProphet[43] and Percolator[46], have been developed to solve this problem.

1.3.2.3.1 Target-Decoy Approach

In 2007, Elias and Gygi[45] developed a simple strategy called target-decoy approach to measure the error content in a search result. Its ease of use quickly gained popularity in the proteomics field. Figure 1.9 illustrates an workflow of the target-decoy approach. In the target-decoy approach, the target database is the normal proteome database, and the decoy database is a randomized version of the target database. Any PSM from the decoy search that passes the selected score threshold is deemed as a false positive. Based on the number of estimated false positives, the global FDR can be readily estimated using the equation in Figure 1.9. Since the underlying assumption of this strategy is that the score distribution of incorrect matches to target sequences is identical to that of matches to decoy ones, the choice of a decoy database becomes critical to the accuracy of the FDR estimation. Several studies[45, 47-49] have been done on this subject and the most common choice of a decoy database is the reversed version of the target database due to its unchanged amino acid composition and identical peptide length distribution. However, it is worth mentioning that target-decoy approach is not as universal as it seems. It imposes some restrictions on the MS/MS matching algorithms, such as whether the algorithm obtains target and decoy matches equally. Those related issues were well discussed in a recent report by Gupta *et al.*[47] In Chapter 4 of this thesis work, we also attempted to examine the applicability of the target-decoy approach.

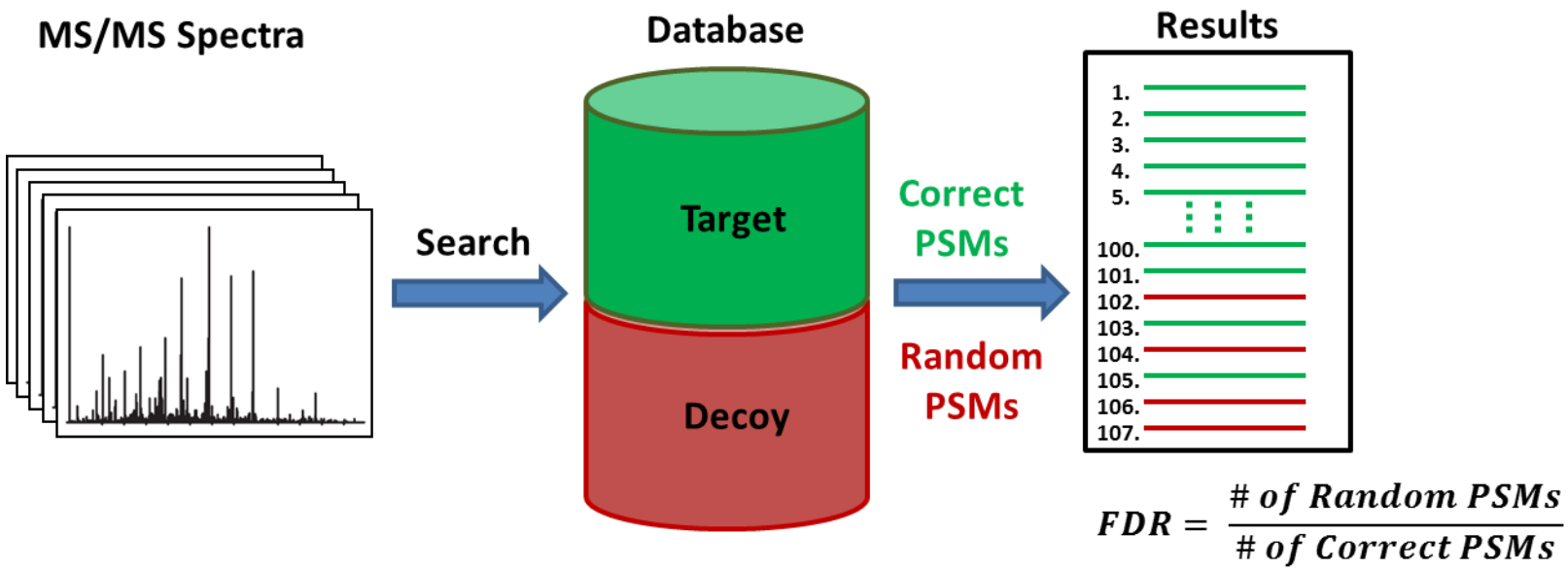

Figure 1.9 Exemplary workflow of using target-decoy approach to estimate FDRs of search results. In this example, the numbers of random and correct PSMs are 4 and 103, respectively. Therefore the estimated FDR is 3.9%.

1.3.2.3.2 PeptideProphet

Before the emergence of the target-decoy strategy, Keller *et al.*[50] had developed a sophisticated algorithm called PeptideProphet to re-evaluate match scores of search results and assign probabilities to PSMs. It takes advantage of the bimodal distribution formed by the discriminant scores of correct and incorrect PSMs in the histogram, uses an expectation-maximization algorithm to resolve the overlapped portion of the distribution and calculate probability of being correct for each PSM as well as the global FDR of the search result at a given threshold (see Figure 1.10). In the early version of PeptideProphet[50], it classifies the correct and incorrect PSMs in an unsupervised fashion (i.e., without having either correct or incorrect PSMs as training sets). Later on, a semi-supervised version of PeptideProphet was developed with the advent of the target-decoy approach.[51] It uses decoy PSMs as the negative training set to help locate the distribution of false PSMs in the target result and thus improve the accuracy of calculated probabilities. Furthermore, another improvement[52] in the PeptideProphet algorithm was made to replace the fixed linear discriminant function with an adaptive one, allowing the algorithm to use more than one PSM for the identification of the best scoring peptide. It has been demonstrated multiple times that PeptideProphet (any version) can significantly improve the number of PSMs of the original search results (e.g., SEQUEST) while maintaining low global FDRs. By using PeptideProphet, not only can one estimate the global FDR of the search result, the probability of each individual PSM being correct is also assigned to measure its reliability.

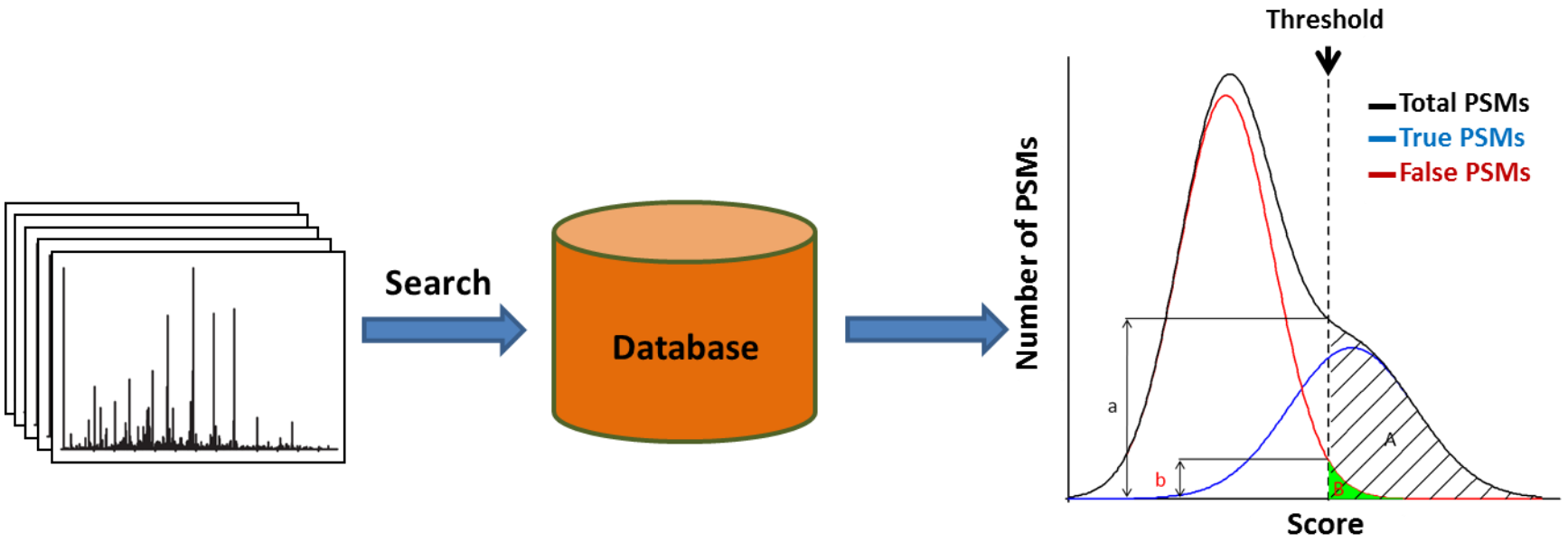

Figure 1.10 Illustration of PeptideProphet strategy. The global FDR can be calculated according to the resolved bimodal distribution (FDR = B/A). The probability of a PSM being correct is defined as 1 – b/a.

1.3.2.3.3 Percolator

Percolator[46], on the other hand, adopts a different machine learning approach. It utilizes target-decoy search results by extracting a vector of features that are related to the quality of the match (e.g., mass error and PSM score) from both target and decoy PSMs. Next, by applying the target-decoy approach on the result, the global FDR is calculated and the target results are then filtered with a fixed FDR (e.g., 1%). This subset of the target results is deemed as the positive training set, while the entire decoy PSMs are treated as the negative training set. Those two data sets are then used for training a support vector machine (SVM), a classification algorithm that analyzes data and recognizes patterns. The learnt classifier is then applied to all the target/decoy PSMs, followed by a new round of FDR calculation, filtering and SVM training as before (Figure 1.11). After several iterations (e.g., 10), the system converges and generates a robust classifier capable of calculating both the probability of each PSM being random (posterior error probability, PEP value) and its associated q-value. Q-value can be understood as the minimum global FDR at which a PSM is accepted. It is basically an extension of global FDR to individual identifications. Studies have shown that the Percolator program can significantly improve the sensitivity of peptide and protein identification for original search engines (e.g., SEQUEST[46] and Mascot[53]). It also showed superior performance than PeptideProphet in some cases.[46] Matrix Science has officially adopted this strategy in its newer version of Mascot server. In Chapter 6, an attempt has been made to interface Percolator program with X!Tandem search engine to enhance the performance of X!Tandem.

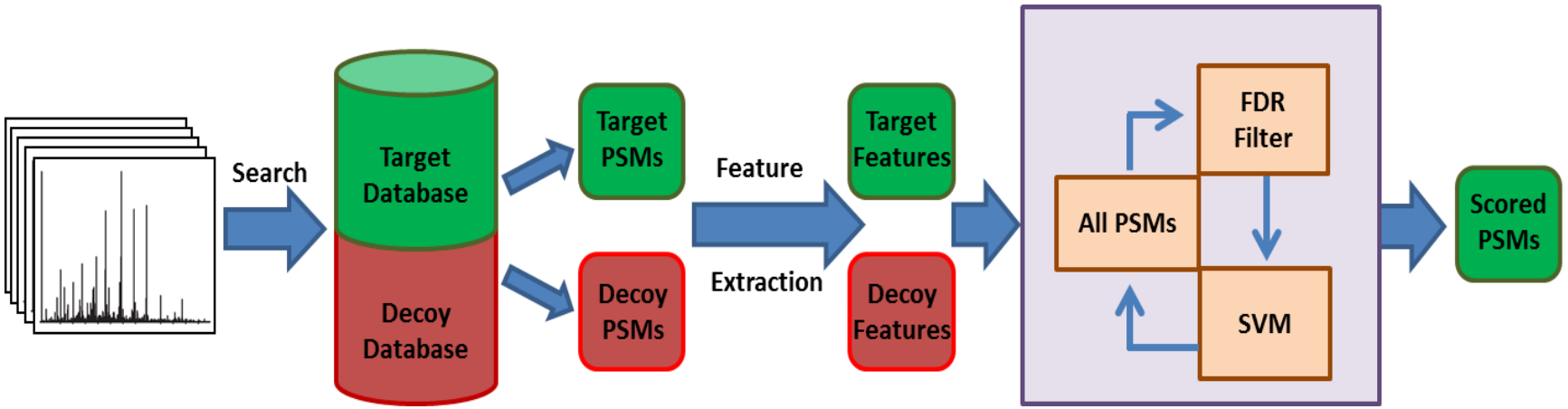

Figure 1.11 Illustration of the Percolator workflow. Adapted from Brosch et al.[41]

With all the statistical evaluation strategies at hand, the proteomic field is still somewhat divided on the best way to provide statistical assessment for search results. Some prefer to compute the error estimates for the entire collection of PSMs (e.g., global FDR and q-value), while some prefer to measure the reliability of each individual PSM (e.g., Mascot ion score, X!Tandem E-value, PeptideProphet probability and Percolator PEP value). Arguments for either can surely be made. In this thesis work, statistical assessment on both individual and global levels is provided to ensure both the reliability of each PSM and low error content in the entire search result.

1.3.3 Spectral Searching

Sequence database searching strategy has provided the proteomics community a great service in terms of correlating MS/MS spectra with peptide sequences. Throughout the past decade or so, an enormous amount of PSMs have been collected. In order to make use of these data, a spectral searching strategy is introduced in the proteomics field. Unlike sequence database searching, where an experimental MS/MS spectrum is compared with a series of theoretical fragmentation patterns to find the best match, in spectral searching the comparison is restricted to pre-identified MS/MS spectra. It turns the peptide spectrum correlation problem into a spectral matching exercise. Due to the reduced search space and more accurate usage of fragment intensities in the spectrum, spectral searching strategy has been demonstrated with much faster speed and higher sensitivity.[54-57]

In fact, spectral searching is not a new concept in the mass spectrometry field. It was first implemented in the mass spectrometric analysis of small molecules.[58-60] The widely used NIST/NIH/EPA mass spectral library developed by the National Institute of Standards and Technology (NIST), contains more than 200,000 mass spectra of small organic molecules (http://www.nist.gov/srd/nist1a.cfm). The very first introduction of spectral searching strategy to proteomics community was in 1998, by Yates and colleagues.[61] Since then, several algorithms have been developed (see Table 1.1) to measure the similarity between experimental spectra and library spectra. Even though they differ in details and score interpretation, most of those algorithms originate from spectral dot product equation, a widely used measurement of similarity between two vectors. However, no study has been carried out to demonstrate which algorithm has the highest sensitivity. It is ascribed to the fact that (1) most of them use custom-made spectral libraries and (2) no standardized statistical evaluation method has been developed for this strategy yet. Since spectral searching strategy is an emerging and promising approach, it is expected that more studies on these subjects will be done in the near future.

Table 1.1 The list of spectral searching programs.

Search Engine	Website
X!Hunter	Web server: http://xhunter.thegpm.org/ Program: ftp://ftp.thegpm.org/projects/xhunter/binaries Library: ftp://ftp.thegpm.org/projects/xhunter/libs
Bibliospec	Program: http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/bibliospec.php Library: http://proteome.gs.washington.edu/software/bibliospec/documentation/
SpectraST	Program: http://sourceforge.net/projects/sashimi/files/ Library : http://www.peptideatlas/speclib/ Web server: http://www.peptideatlas.org/spectrast/
NISTMS Search	http://peptide.nist.gov/
SpecMatching	Upon request

1.4 Spectral Library Construction

Besides a robust searching algorithm, the success of a spectral search also relies heavily on MS/MS libraries. In practice, the typical process of constructing a spectral library from experimental MS/MS raw spectra consists of five stages (see Figure 1.12). First, the spectra are correlated with peptide sequences by sequence search engines (e.g., Mascot). Second, validation (e.g., statistical validation) is performed to filter out unreliable identifications, thus ensuring the quality of spectral library in later stages. Third, the original MS/MS spectra and their corresponding identification information are linked together. By doing so, additional information such as the annotation of peaks in the spectra and experimental condition under which spectra were collected is meticulously preserved. Fourth, spectral processing procedures such as noise reduction and replicate spectra consolidation are needed to construct library spectra, since the sensitivity and speed of spectral search engines is closely related to the richness of spectral information (e.g., number of peaks in the library spectra) and redundancy of the spectral library. While some prefer to generate one "consensus" spectrum for each sequence from merging multiple replicate identifications[56], some choose the best replicate spectral identification as the representative.[55] The fifth and final step is compilation of all the processed spectral identifications into a usable library that meets the format requirement of the spectral search engines. All the related topics have been recently reviewed by Lam.[62]

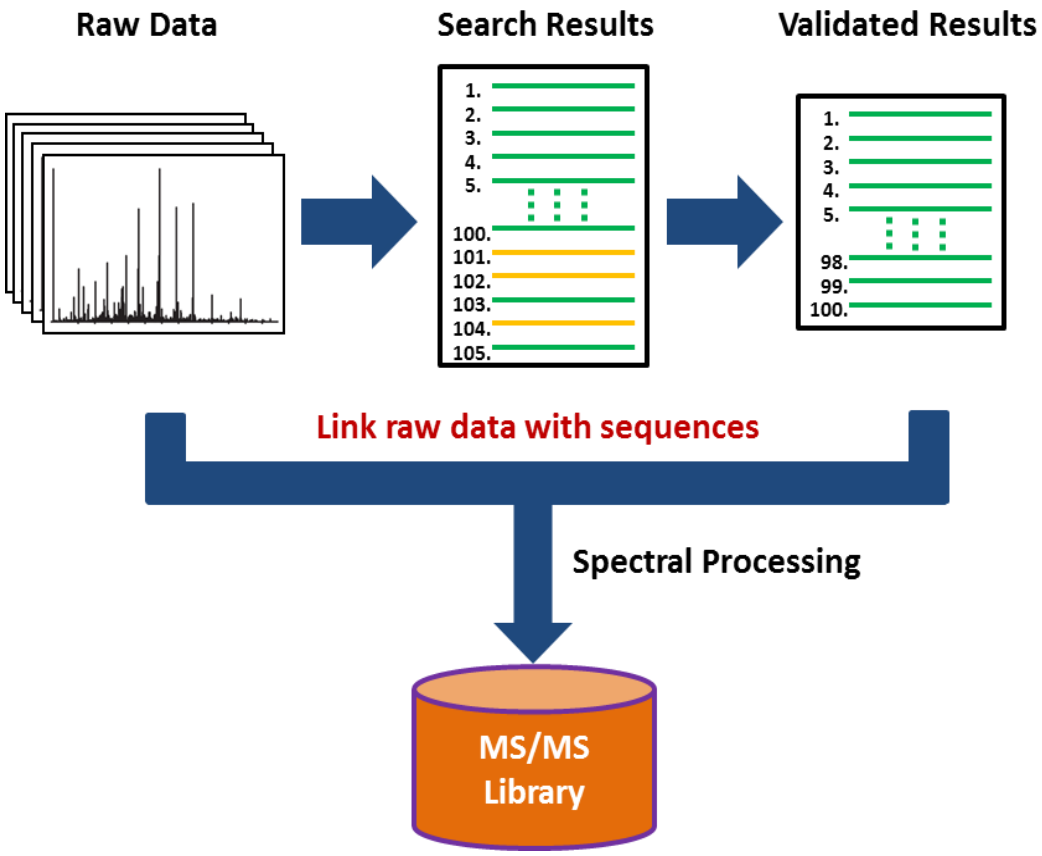

Figure 1.12 The typical procedure of spectral library building, including (1) the collection of MS/MS spectra, (2) correlation of raw spectra with peptide sequences, (3) validation of identifications, (4) spectral processing and (5) compilation of reliable MS/MS libraries.

1.5 Experimental Validation

As mentioned in the previous section, a reliable spectral library is the foundation of spectral searching strategy. During the process of spectral library construction, validation is always needed to ensure that only confident identifications are compiled into the library. Nowadays, most validation methods used in this process are based on a specific statistical model, such as sequence search engine scores and PeptideProphet probability (see section 1.3.2.3). Because

using sequence database searching strategy to elucidate peptide sequences is far from perfect, some researchers sought to use stable isotope labeling to facilitate and validate the spectral interpretation.[63]

Stable isotope labeling strategy introduces defined mass changes on labeled peptides or proteins that can be distinguished by a mass spectrometer. It is widely applied in quantitative proteomics. Three typical labeling strategies are illustrated in Figure 1.13, (1) chemical labeling by isotopically encoded reagents, which are usually a pair of reagents in light and heavy isotope form; (2) incorporation of ^{18}O isotope via peptide bond hydrolysis in ^{18}O enriched water; and (3) metabolic labeling. In the first strategy, labeling could be performed on either protein or peptide level, which results in chemical modification of functional groups on the amino acid residues. The ^{18}O enzymatic labeling incorporates ^{18}O isotope in the peptide sequence at the same time as the digestion, resulting in almost no chemical property change for the peptides. In the metabolic labeling strategy, proteins are metabolically labeled via cell culture in isotope enriched medium (e.g., ^{13}C-enriched glucose). Because the peptide incorporated with heavy isotope(s) will have higher mass but nearly identical chemical behavior as the same sequence incorporated with light isotope(s), by comparing their intensity in the mass spectrometric analysis, relative quantitation can be achieved. During the past decade or so, numerous labeling methods have been developed and applied to various biological systems. A recent review has summarized and discussed the strength and weakness of all different kinds of stable isotope labeling methods for quantitative proteomic studies.[64]

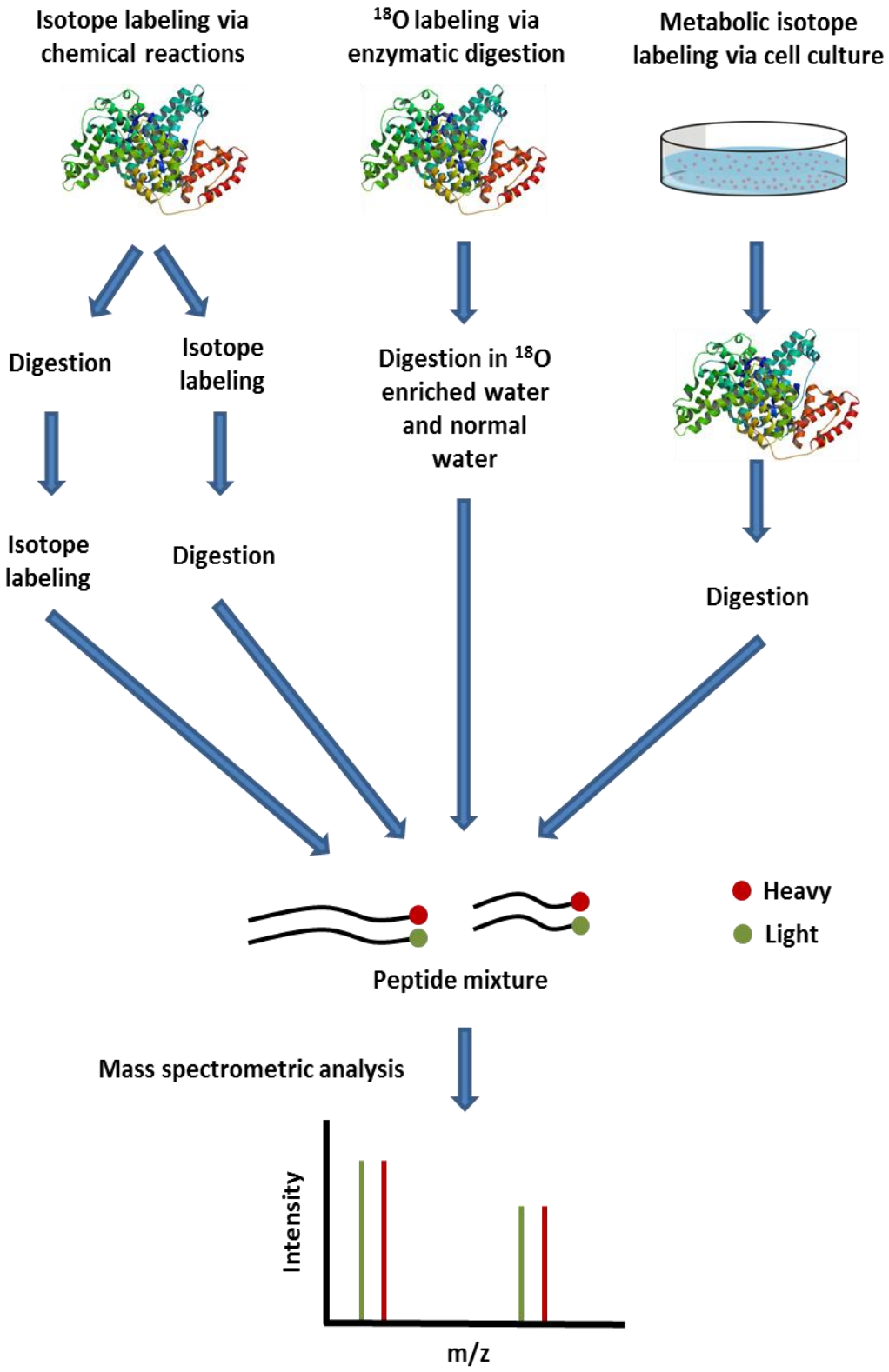

Figure 1.13 Three stable isotope labeling strategies in comparative proteomics.

Several authors have recognized that stable isotope labeling could be also used to facilitate and validate the interpretation of MS/MS spectra for peptide identification.[63, 65-68] The concept is fairly straightforward. When overlaying the MS/MS spectra of the same peptide sequence labeled with heavy and light isotope(s), true peptide fragment ions will show up as doublet peaks while background peaks or true peptide fragment ions not containing the isotopic group will remain singlet (see Figure 1.14). In both *de novo* sequencing[63, 69, 70] and sequence database searching[65, 68] studies, stable isotope labeling has been proven very effective in elucidation and validation of peptide sequence assignments. Specifically, metabolic and ^{18}O-enzymatic labeling methods are often the method of choice. In both strategies, all the functional groups of amino acid residues in a peptide sequence remain unmodified after the labeling, which presents an opportunity to understand the fragmentation pathways of various peptides.

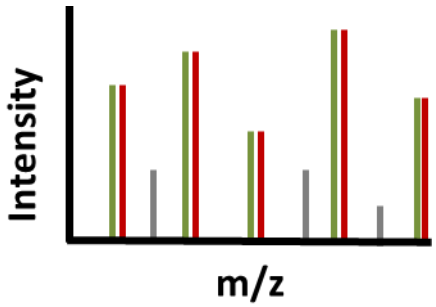

Figure 1.14 Overlaid MS/MS spectrum of the same peptide sequence labeled heavy and light isotope(s). True peptide fragment ions show up as doublet peaks (red and green). Background peaks are shown as grey.

Gu and co-workers[71] introduced a metabolic labeling method by incorporating lysine-d_4 (4 deuterium atoms) into proteins during cell culture. The lysine-d_4 labeled protein then was digested by endoproteinase Lys-C, resulting in

all the peptides having a C-terminal lysine except for the ones from the original C-termini of proteins. By comparing the MS/MS spectra of the unlabeled and labeled peptide of the same sequence, in which the C-terminal fragment ions showed as doublets with 4.025-Da mass difference, the C-terminal fragment assignment was unambiguously validated. Hence it gave rise to highly confident peptide spectral identifications.

Similarly, Zhong *et al.*[68] used a ^{15}N-metabolic labeling method to demonstrate that by simply overlaying the unlabeled and ^{15}N-labeled MS/MS spectra of the same peptide sequence, experimental evidence can be provided to validate protein identification results generated by the sequence database search method. A unique aspect of ^{15}N-labeling is that the fragment ions composed of different amino acid sequences have non-uniform mass shifts from their corresponding unlabeled fragment ions, facilitating the identification of correct assignment from other false matches.

However, the metabolic labeling method has its drawbacks as well. It tends to be relatively costly as lengthy growth periods are required to ensure thorough isotope incorporation. Another major limitation of metabolic labeling methods is that samples need to be cultured on a specific isotope enriched medium. Therefore, not all biological systems are eligible for metabolic labeling. Samples such as human blood and plasma, or samples arising from other naturally occurring sources cannot be metabolically labeled. As a complementary approach to metabolic labeling, ^{18}O-labeling is less expensive and applicable to almost all

biological systems.[72] Therefore, it is also frequently applied to validate peptide-sequence assignments.

Various protocols[73-75] have been developed to optimize the conditions for incorporating ^{18}O and minimize $H_2{}^{18}O$ consumption. Meanwhile, to minimize the back exchange of ^{18}O to ^{16}O after labeling, researchers have proposed numerous strategies, such as using prohibitors[76] or heat[77] to reduce the enzyme activity and using immobilized trypsin[76] instead of solution based trypsin.

In trypsin-mediated ^{18}O-exchange, two ^{18}O atoms are substituted for the two C-terminal ^{16}O atoms for all the tryptic peptides containing C-terminal lysine or arginine. Comparison of the fragmentation patterns of ^{18}O-labeled and unlabeled peptides of the same sequence (4.0085-Da mass difference) distinguishes b- and y- ions. The C-terminal fragments (y ions) appear as doublet peaks in the MS/MS spectrum, but the N-terminal fragments (b ions) display as singlet peaks. Because of the fragmentation preference of tryptic peptides, relatively higher basicity of lysine and arginine at the C-terminus of tryptic peptides gives rise to a dominant y- ion series in the MS/MS spectrum.

This strategy has been productively applied to *de novo* sequencing[78-80], successfully elucidating the sequence information and unambiguously annotating fragment ions in MS/MS spectra. In sequence database search, ^{18}O-labeling has also been employed to validate peptide-spectrum assignments and even remove noise in MS/MS spectra.[81, 82] In 2009, Volchenboum *et al.*[65] created data sets by performing LC-MS/MS analysis on a mixture with ^{18}O-labeled and unlabeled

peptides in equal amounts and subsequently developed a set of software tools to provide rapid and automatic validation of peptide assignments by Mascot. It successfully demonstrated that by employing the ^{18}O-labeling strategy, many true identifications deemed as insignificant by Mascot can be re-captured, improving the sensitivity and specificity of the sequence database searching strategy.

Compared to statistical validation, where probability of being correct or random is assigned to peptide-spectrum matches, experimental validation offers concrete evidence and definitive conclusions. However, extra steps in the sample preparation procedure make the experimental validation strategy more expensive, more time-consuming, less robust and consequently less popular. Nonetheless, its high accuracy of detecting true identifications can be still quite useful for constructing reliable MS/MS libraries and evaluating the performance of search engines. In this thesis work, both statistical evaluation and experimental validation were employed.

1.6 Scope of the Thesis

The main objective of this work was to develop a sensitive spectral searching strategy for shotgun proteomic studies. In Chapter 2, a mass spectrometric method was developed for the analysis of proteome from thousands of cancer cells. In Chapter 3, using *E. coli* K12 cell lysates as the model system and ^{15}N-metabolic labeling as the experimental validation strategy, peptide-

spectrum matches by Mascot were validated and further used to construct a reliable MS/MS spectral library for spectral searching strategy. A spectral searching algorithm was also developed to utilize the highly confident spectral library. In Chapter 4, in combination with the ^{18}O-labeling method, an inclusion strategy was developed to experimentally validate the peptide matches from human cell lysates by sequence search engines. Using those experimentally validated matches, the performance of commonly used statistical tools was evaluated. In Chapter 5, by examining the validated matches from Chapter 4, strategies were proposed to detect single-hit proteins (proteins identified by only one peptide sequence) with high confidence. In Chapter 6, by using the experimentally validated data set from chapter 3, X!Tandem, an open source sequence search engine, was successfully coupled with Percolator, one of the most powerful statistical validation tools, to provide sensitive and accurate peptide identifications. Finally, conclusions and future work involving construction of spectral libraries and application of spectral searching strategy are described in Chapter 7.

1.7 Literature Cited

(1) Anderson, N. L.; Anderson, N. G. *Electrophoresis* **1998**, *19*, 1853-1861.

(2) Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O.; Sanchez, J.-C.; Yan, J. X.; Gooley, A. A.; Hughes, G.; Humphery-Smith, I.; Williams, K. L.; Hochstrasser, D. F. *Nat. Biotechnol.* **1996**, *14*, 61-65.

(3) Blackstock, W. P.; Weir, M. P. *Trends Biotechnol.* **1999**, *17*, 121-127.

(4) Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837-846.

(5) Lill, J. R.; Ingle, E. S.; Liu, P. S.; Pham, V.; Sandoval, W. N. *Mass Spectrom. Rev.* **2007**, *26*, 657-671.

(6) Simpson, D. M.; Beynon, R. J. *J. Proteome Res.* **2009**, *9*, 444-450.

(7) Wright, A. P. H.; Bruns, M.; Hartley, B. S. *Yeast* **1989**, *5*, 51-53.

(8) Speers, A. E.; Wu, C. C. *Chem. Rev.* **2007**, *107*, 3687-3714.

(9) Kashino, Y. *J. Chromatogr. B* **2003**, *797*, 191-216.

(10) Carboni, L.; Piubelli, C.; Righetti, P.; Jansson, B.; Domenici, E. *Electrophoresis* **2002**, *23*, 4132-4141.

(11) Jones, M. *Int. J. Pharm.* **1999**, *177*, 137-159.

(12) Schägger, H.; von Jagow, G. *Anal. Biochem.* **1987**, *166*, 368-379.

(13) Eschelbach, J. W.; Jorgenson, J. W. *Ana. Chem.* **2006**, *78*, 1697-1706.

(14) Urbas, L.; Brne, P.; Gabor, B.; Barut, M.; Strlič, M.; Petrič, T. Č.; Štrancar, A. *J. Chromatogr. A* **2009**, *1216*, 2689-2694.

(15) Fournier, M. L.; Gilmore, J. M.; Martin-Brown, S. A.; Washburn, M. P. *Chem. Rev.* **2007**, *107*, 3654-3686.

(16) Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. *J. Am. Chem. Soc.* **1999**, *121*, 806-812.

(17) Blackler, A. R.; Speers, A. E.; Ladinsky, M. S.; Wu, C. C. *J.Proteome Res.* **2008**, *7*, 3028-3034.

(18) Zhong, H.; Zhang, Y.; Wen, Z.; Li, L. *Nat. Biotechnol.* **2004**, *22*, 1291-1296.

(19) Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* **2009**, *81*, 1049-1060.

(20) Fenn, J.; Mann, M.; Meng, C.; Wong, S.; Whitehouse, C. *Science* **1989**, *246*, 64-71.

(21) Iribarne, J. V.; Thomson, B. A. *J. Chem. Physics* **1976**, *64*, 2287-2294.

(22) Kebarle, P. *J. Mass Spectrom.* **2000**, *35*, 804-817.

(23) Kebarle, P.; Peschke, M. *Anal. Chim. Acta* **2000**, *406*, 11-35.

(24) Karas, M.; Bachmann, D.; Hillenkamp, F. *Anal. Chem.* **1985**, *57*, 2935-2939.

(25) Mamyrin, B. A. *Int. J.Mass Spectrom.* **2001**, *206*, 251-266.

(26) Bogdanov, B.; Smith, R. D. *Mass Spectrom. Rev.* **2005**, *24*, 168-200.

(27) Makarov, A. *Anal. Chem.* **2000**, *72*, 1156-1162.

(28) Dass, C. *Fundamentals of Contemporary Mass Spectrometry*; Wiley Hoboken, New Jersey, 2007.

(29) Wang, N.; Li, L. *Anal. Chem.* **2008**, *80*, 4696-4710.

(30) Paizs, B.; Suhai, S. *Mass Spectrom. Rev.* **2005**, *24*, 508-548.

(31) Bensadek, D.; Monigatti, F.; Steen, J. A. J.; Steen, H. *Int. J. Mass Spectrom.* **2007**, *268*, 181-189.

(32) Pandey, A.; Andersen, J. S.; Mann, M. *Sci. STKE* **2000**, *2000*, pl1-.

(33) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci.* **1996**, *93*, 14440-14445.

(34) Wang, W.; Sun, J.; Nimtz, M.; Deckwer, W.-D.; Zeng, A.-P. *Proteome Sci.* **2003**, *1*, 6.

(35) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327-342.

(36) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. *Mol. Cell. Proteomics* **2007**, *6*, 1638-1655.

(37) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(38) Fenyoe, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.

(39) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426-1436.

(40) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y.-K.; Yoo, J.-S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M. *Proteomics* **2005**, *5*, 3226-3245.

(41) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. *Mol. Cell. Proteomics* **2008**, *7*, 962-970.

(42) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 47-50.

(43) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(44) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531-533.

(45) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(46) Kaell, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

(47) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111-1120.

(48) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. *Proteomics* **2003**, *3*, 1454-1463.

(49) Klammer, A. A.; MacCoss, M. J. *J. Proteome Res.* **2006**, *5*, 695-700.

(50) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(51) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 254-265.

(52) Ding, Y.; Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 4878-4889.

(53) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.

(54) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. *J. Proteome Res.* **2006**, *5*, 1843-1849.

(55) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. *Anal. Chem.* **2006**, *78*, 5678-5684.

(56) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *Proteomics* **2007**, *7*, 655-667.

(57) Xu, M.; Li, L. *J. Proteome Res.* **2011**, *10*, 3632-3641.

(58) Stein, S.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859-866.

(59) Domokos, L.; Henneberg, D.; Weimann, B. *Anal. Chim. Acta* **1984**, *165*, 61-74.

(60) Owens, K. G. *Applied Spectroscopy Rev.* **1992**, *27*, 1-49.

(61) Yates, J. R., III; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557-3565.

(62) Lam, H. *Mol. Cell. Proteomics* **2011**, *10*, 008565/008561-008565/008510.

(63) Gu, S.; Pan, S.; Bradbury, E. M.; Chen, X. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 1-7.

(64) Kline, K. G.; Sussman, M. R. *Ann. Rev. of Biophy.* **2010**, *39*, 291-308.

(65) Volchenboum, S. L.; Kristjansdottir, K.; Wolfgeher, D.; Kron, S. J. *Mol. Cell. Proteomics* **2009**, *8*, 2011-2022.

(66) Takao, T.; Hori, H.; Okamoto, K.; Harada, A.; Kamachi, M.; Shimonishi, Y. *Rapid Commun. Mass Spectrom.* **1991**, *5*, 312-315.

(67) Gu, S.; Chen, X. *Analyst* **2005**, *130*, 1225-1231.

(68) Zhong, H.; Marcus, S. L.; Li, L. *J. Proteome Res.* **2004**, *3*, 1155-1163.

(69) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015-1024.

(70) Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002**, *20*, 163-170.

(71) Gu, S.; Pan, S.; Bradbury, E. M.; Chen, X. *Anal. Chem.* **2002**, *74*, 5774-5785.

(72) Ye, X.; Luke, B.; Andresson, T.; Blonder, J. *Brief. Func. Genomic. Proteomics* **2009**, *8*, 136-144.

(73) Yao, X.; Afonso, C.; Fenselau, C. *J. Proteome Res.* **2002**, *2*, 147-152.

(74) Miyagi, M.; Rao, K. C. S. *Mass Spectrom. Rev.* **2007**, *26*, 121-136.

(75) Hajkova, D.; Rao, K. C. S.; Miyagi, M. *J. Proteome Res.* **2006**, *5*, 1667-1673.

(76) Sevinsky, J. R.; Brown, K. J.; Cargile, B. J.; Bundy, J. L.; Stephenson, J. L. *Anal. Chem.* **2007**, *79*, 2158-2162.

(77) Storms, H. F.; van der Heijden, R.; Tjaden, U. R.; van der Greef, J. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 3491-3497.

(78) Back, J. W.; Notenboom, V.; de Koning, L. J.; Muijsers, A. O.; Sixma, T. K.; de Koster, C. G.; de Jong, L. *Anal. Chem.* **2002**, *74*, 4417-4422.

(79) Qin, J.; Herring, C. J.; Zhang, X. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 209-216.

(80) Lee, Y. H.; Han, H.; Chang, S.-B.; Lee, S.-W. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 3019-3027.

(81) Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A. *Proc. Natl. Acad. Sci.* **2007**, *104*, 6140-6145.

(82) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *Mol. Cell. Proteomics* **2006**, *5*, 935-948.

Chapter 2

Development of a Shotgun Method Based on Liquid Chromatography Quadrupole Time-of-flight Mass Spectrometry for Proteome Analysis of 500 to 5000 Cancer Cells[*]

2.1 Introduction

Current mass spectrometric technology can identify thousands of proteins from a proteome sample and has become a powerful and popular tool for mapping the entire proteome. This large scale proteome profiling work is generally carried out using hundreds of micrograms or milligrams of starting materials. To produce this quantity of sample, millions or even billions of cells are used. For cultured cells, the enormous demand of cells is usually not a major issue. However, in many other studies, the number of cells available for proteome analysis can be quite limited. For example, in a tissue sample containing both normal and transformed (e.g., cancer) cells, the number of cancerous cells may be very limited.[1] This is particularly true for tissue samples from patients at an early stage of cancer development.[2, 3] Another example is the characterization of the proteome from a small number of circulating cancerous cells in a blood sample of a patient with early sign of a tumor in a specific organ.[4]

[*] A version of this chapter has been published as Nan Wang, Mingguo Xu, Peng Wang and Liang Li, **2010**, "Development of Mass Spectrometry-Based Shotgun Method for Proteome Analysis of 500 to 5000 Cancer Cells", *Anal. Chem.* 82, 2262-2271. My contribution included sample preparation, method development and data analysis.

Our research goal is to develop new techniques that can generate as large a proteome coverage as possible from a small number of cells. Our initial target is to analyze the proteome of about 1000 cells. Adequate coverage of the proteome from this number of cells may lead to several important applications. For example, 1000 cells may be collected from a patient blood containing rare circulating cancerous cells from an early stage of metastasis of a solid tumor.[5-9] Analyzing the proteome of these cells may be used as a fingerprint for diagnosis or prognosis of a cancer. Another example is that about 1000 cells may be procured from a tissue section using laser capture microdissection (LCM) within a couple of hours. Analyzing these cells may assist in identifying specific protein markers for disease diagnosis. The ultimate goal of this research is to analyze the single cell proteome.[10] Unfortunately, this is a huge challenge at this moment for mass spectrometry based technologies due to limited sensitivity. Developing and applying techniques for analyzing the proteome of thousands of cells is a more realistic goal. However, very few studies of proteome analysis from a few thousands of cells have been reported.[11-15]

In this chapter, a shotgun proteome analysis method is described for analyzing proteomes of MCF-7 cells ranging from 500 to 5000 cells. MCF-7 cells, derived from breast cancer, are representative of many different types of cancerous cells in terms of size and proteome complexity. Thus, the method developed from analyzing MCF-7 cells should be applicable to other cancerous cells. The performance of this method in terms of the numbers of peptides and proteins identifiable from small numbers of cells is reported. This method is then

applied to a model system where a small number of MCF-7 cells are added to human blood to mimic a patient blood sample containing cancerous cells. These cells are captured by the combination of antibody attachment to the cells and flow cytometry for cell sorting. The captured cells are analyzed by the shotgun proteomic method.

2.2 Experimental

2.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), iodoacetamide (IAA), trifluoroacetic acid (TFA), sodium bicarbonate were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, HPLC grade formic acid, LC-MS grade water, acetone, and acetonitrile (ACN) were purchased from Fisher Scientific Canada (Edmonton, Canada). The BCA assay kit was obtained from Pierce (Rockford, IL).

2.2.2 Cell Preparation

Figure 2.1 shows the overall workflow for the shotgun method used to analyze a small number of cells. The MCF-7 breast cancer cells (ATCC® number: HTB-22™) were cultured in 15 cm diameter plates at 37 ℃ in DMEM Gibco medium supplemented with 10% fetal bovine serum. The plates were then washed twice with ice-cold 25 mL PBS^{++} buffer (0.68 mM $CaCl_2$, 0.5 mM $MgCl_2$, 1.4 mM KH_2PO_4, 4.3 mM Na_2HPO_4, 2.7 mM KCl, and 137 mM NaCl). The cells

were harvested by scraping from the plates into the PBS^{++} buffer and centrifugation at 100 *g* for 8 min at 4 ℃. The cell numbers were first roughly counted by an Axiovert 25 hemocytometer (Carl Zeiss, Inc. Minneapolis, MN).

The fresh whole blood provided by a healthy donor was first diluted in PBS buffer (1.4 mM NaCl, 0.27 mM KCl, 1 mM Na_2HPO_4, 0.18 mM KH_2PO_4, pH 7.4) in a 1:10 ratio (v:v). The MCF-7 human breast cancer cells (2 million) were then spiked into the diluted blood solution. Density separation was then conducted to remove red blood cells by using Ficoll-Hypaque (GE Healthcare) (See Figure 2.2). In brief, 10 mL diluted blood sample containing MCF-7 cells blood was slowly added into 4 mL Ficoll solution. The solution was spun down at 2000 rpm, 4 ℃ for 20 min. Considering the density of MCF-7 cells, the cancer cells preferentially aggregated with peripheral blood leukocytes (PBL) at the layer called buffy coat after centrifugation. The buffy coat was isolated, washed and re-suspended in PBS buffer. Afterwards, the cell mixture, was incubated with a FITC–conjugated mouse anti–human HEA antibody (Miltenyi Biotec number: 130-080-301) in a 1:100 (v:v) ratio on ice for 15 min. Therefore, most MCF-7 cells were fluorescently stained, while PBL were not.

Both the unstained MCF-7 cells and the stained cell mixtures were introduced into the flow cytometer (Beckman Coulter EPICS Altra) for counting, according to the cell size and their fluorescence response. Then 500, 1000, 2500 or 5000 MCF-7 cells were collected into 0.6 mL low retention micro-centrifuge vials (Fisher Scientific).

2.2.3 Protein Extraction and Digestion

The cells in each vial were mixed with 5 to 10 µL Nonidet-P40 (NP40) lysis buffer (1%) and sonicated in ice-water ultrasonic bath for 5 min. The protein solutions were then reduced with 20 mM (0.4 to 0.75 µL) dithiothreitol (DTT) and alkylated with the same volume of 40 mM iodoacetamide. Acetone (precooled to -80 ^{o}C) was added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The solution was then incubated at -20 ℃ for 60 min and centrifuged at 14 000 rpm for 10 min. The supernatant was decanted. The pellet was carefully washed once using cold acetone to ensure the efficient removal of NP40 detergent (See Figure 2.1). The residual acetone was evaporated at ambient temperature. Then 50 mM ammonium bicarbonate was used to sufficiently re-dissolve the pellet in the vial. Trypsin digestion was then carried out in a final enzyme concentration of 8 ng/µl (5 to 20 uL) at 37 ^{o}C for 4 hours.

2.2.4 Peptide Desalting and Quantification by RPLC

The desalting and quantification setup consisted of an Agilent 1100 HPLC system (Palo Alto, CA) with a UV detector. The desalting of tryptic peptides was performed on a 4.6 mm × 50 mm Polaris C18 A column with 3 µm particle and 300 Å pore (Varian, CA). After loading all the digests of each sample, the column was flushed at 1 mL / min with 97.5% mobile phase A (0.1% TFA in water) for 3 min and then 85% of mobile phase B (0.1% TFA in acetonitrile) for 5 min to ensure the complete elution of the peptide fractions from the column.

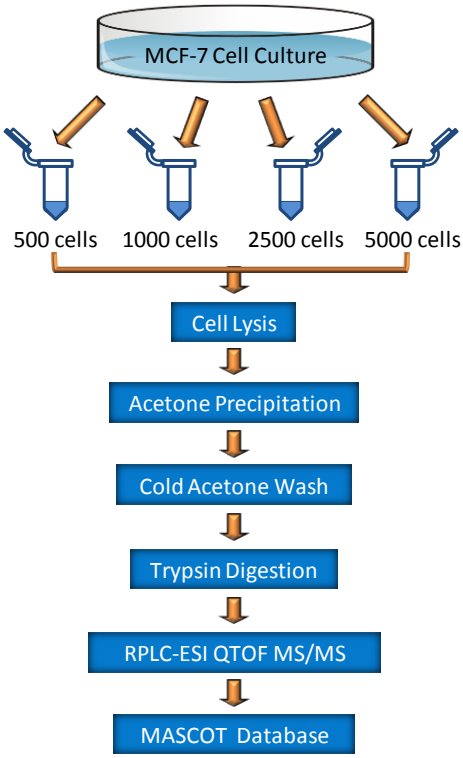

Figure 2.1 Workflow for both method development and application.

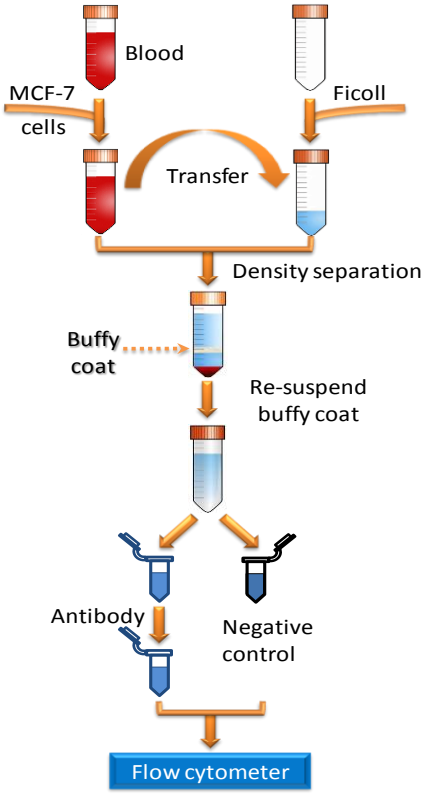

Figure 2.2 Workflow for the enrichment of MCF-7 cells in a blood sample.

2.2.5 LC-ESI QTOF MS and MS/MS Analysis

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, the desalted and quantified digests were concentrated using a SpeedVac (Thermo Savant, Milford, MA) to ~1 µl and reconstituted to a specific concentration using 0.1% formic acid. Then the intended amount of digest solution was injected onto a 75 µm × 100 mm Atlantis dC18 column (Waters, Milford, MA). For the digests from 500 and 1000 cells, multiple injections were applied for each sample to make sure the maximum amount of peptides was loaded. Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using an optimal gradient elution ranging from 90 min to 270 min in length and electrosprayed into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A survey MS scan was acquired from m/z 350-1600 for 0.8 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 0.8 s each. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants (i.e., lock-mass), was infused at a flow rate of 300 nL/min, and a 1 s MS scan was acquired every 1 min throughout the run.

2.2.6 Protein Database Search

Raw LC-ESI data were lock-mass corrected, deconvoluted, and converted to peak list files by using ProteinLynx Global Server 2.2.5 (Waters). Peptide sequences were identified via automated database searching of peak list files

using the Mascot search program (version 1.8). Database searching was restricted to *Homo sapiens* (human) in the SWISSPROT database (October 4, 2007) and 17317 entries were searched. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, 30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, acetyl (Protein), oxidation (M), pyro-Glu (N-term Q) and pyro-Glu (N-term E). The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, Mascot threshold score for identity, calculated molecular mass of the peptide, and the difference between the experimental and calculated masses were extracted to Excel files using an in-house program. All the identified peptides with scores lower than the Mascot threshold score for identity at a confidence level of 95% were then removed from the protein list. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list.

Because of the small data set generated from the proteome analysis of a few cells, accurate analysis of the false discovery rate (FDR) is difficult. The commonly used target-decoy search strategy is best suited for analyzing a large data set.[16-18] To ensure data quality, many of the matched MS/MS spectra with peptide sequences were manually validated. Specifically, peptide matches with a matching score less than 10 points above the Mascot identity threshold were manually analyzed. The peptide match was considered as positive identification if

the fragment ions contained more than five isotopically resolved y-, b-, or a-ions and the major fragment ion peaks with high intensity (i.e., peak intensity of >30% in a normalized spectrum). Most of the high intensity fragment ions (i.e., top 5) must also belong to y-, b-, or a-ions, not internal fragment ions. Peptide matches which failed to meet these criteria were removed from the protein lists. Typically, this manual validation process eliminated about 3% of the low score matches. A protein was considered to be identified even if a single peptide match was found.

2.3 Results and Discussion

Shotgun proteome analysis is a relatively sensitive technique, compared to other methods such as gel-based proteome analysis.[19] For example, about 1 μg of a cell extract digest injected to LC-ESI MS/MS can result in the identification of about 300 to 800 proteins, depending on the complexity of the sample. In the shotgun method, the sample workup process includes cell lysis, protein extraction, protein digestion and injection of peptides into the LC-MS/MS system for analysis. Any one of the steps can potentially involve the loss of some proteins. In working with a large quantity of samples, this sample loss may not be very significant so long as the sample loss is not biased towards a particular group of proteins. If a bias (i.e., selective sample loss) does occur, that group of proteins will be under-represented in the final results. If the sample loss is un-biased, as long as there are sufficient amounts of peptides for LC-MS/MS analysis (e.g., 1 μg per injection), the same proteome coverage would be expected. However, in

handling small numbers of cells, sample loss of any type can be detrimental to the proteome coverage. The reason is that the amount of sample generated from a small number of cells will be limited and it will often not meet the optimal sample amount required for peptide sequencing in LC-MS/MS (e.g., < 1 μg). In a recent report, Wang et al. have shown that the amount of sample injection is very important in determining the outcome of peptide and protein identification.[16] Injection of a smaller amount of sample results in a lower number of peptides and proteins identified. For the nano-LC QTOF MS platform used in this work, the optimal amount of peptides for injection is about 1 μg and exceeding this amount does not result in a significant increase in peptide and protein numbers.

With the above considerations in mind, a sample analysis protocol was developed as shown in Figure 2.1. In our work, instead of taking an aliquot from a stock solution containing a large number of cells to make a sample of a small number of cells, the cultured MCF-7 cells were sorted into tubes containing different numbers of cells using a flow cytometer. In searching for a suitable sample preparation protocol to handle small numbers of cells, a LC-UV technique was used to measure the amount of peptides produced by individual protocols tested and to compare the peptide amounts to determine which protocol yielded the highest peptide amount. Three surfactant-based methods using sodium dodecyl sulfate (SDS), acid labile surfactant (ALS) from Waters or a cell lysis solution containing NP-40 detergent as well as two special reagents using Tris buffer or trifluoroethanol (TFE) were examined for their performance in cell lysis and downstream sample workup. The surfactant-based methods are widely used

for efficient cell lysis in the proteome analysis work involving large numbers of cells.[20, 21] In the case of SDS, after cell lysis and trypsin digestion, SDS had to be removed by a strong cation exchange column to reduce its interference with LC-ESI MS. For ALS, the tryptic digest was acidified to degrade ALS and the hydrophobic products were carefully removed, prior to MS analysis. The use of Tris buffer or NP-40 cell lysis solution was straightforward by mixing the solution with the sample with intermittent sonication, followed by trypsin digestion. TFE was used according to the reported protocol. Among the five protocols tested, using NP-40 lysis solution, the average amount (n=3) of peptides from the 5000-cell sample was found to be the highest (1.40 ± 0.12 μg). However, one major problem initially encountered in using this polyethylene glycol based detergent for cell lysis was that, after acetone precipitation of proteins from the lysate, the pellet still contained a small amount of NP-40, causing severe interference in LC-ESI MS analysis of the cell lysate protein digest. To eliminate this interference, the pellet was carefully washed with cold acetone. This simple step was found to be very effective in reducing the NP-40 content to a level that did not cause interference in LC-ESI MS. As Figure 2.1 shows, the cold-acetone washed pellet was dissolved in NH_4HCO_3, followed by trypsin digestion. The tryptic digest was desalted, quantified and then injected into the LC-ESI QTOF instrument for MS/MS sequencing of the peptides.

The amount of peptides produced from a cell lysate was determined using the LC-UV system as described in the reference.[16] The average amount (n=3) of peptides from the 5000-cell sample was found to be 1.40 ± 0.12 μg. And the

average amount of the 2500-cell sample was 0.83 ± 0.12 μg, which is not exactly half of the amount of peptides produced from the 5000-cell sample. But, within the experimental errors, the amount of peptides produced appears to proportionally decrease as the cell number decreases. If this proportionality held true for the 1000- or 500-cell sample, then the amount of peptides produced would be less than 0.28 μg for the 1000-cell sample and 0.14 μg for the 500-cell sample. The lower limit of the UV-LC system used to measure the peptide concentration is about 0.25 μg. An attempt was made to measure the peptide amounts for the 1000- and 500-cell samples and the results were not reliable as they generated UV signals with intensities similar to that of the blank. The failure to quantify the 1000-cell sample suggests that the amount of peptides produced from this sample must be less than 0.25 μg. Thus, sample loss may be more severe for these two samples, compared to the 2500- or 5000-cell sample. This is understandable as the same protocol was applied to these samples and the same amount loss (e.g., via adsorption to the container walls) would result in a greater percentage of sample loss for the 1000- or 500-cell samples. For future work, a simple and accurate quantification method to determine nano-grams of peptides or proteins in each step of the workflow shown in Figure 2.1 should facilitate the optimization process. One approach is to modify the current LC-UV system using a capillary column, instead of a 1 mm column, to shift the linear calibration curve to the nano-gram region.

Besides sample preparation, optimization of the LC-ESI MS runs is also critical in analyzing samples of a few cells. In our work, a trap column was used

to facilitate the peptide loading to the nano-LC QTOF MS instrument. For sample injection, the minimum volume of residual sample required to be present in the sample vial is about 1 μL. In our experiment, after drying the desalted samples, each sample was redissolved to make 11 L of solution by adding 0.1% formic acid from which two injections with each 5 L were carried out. These two injections with 1 μL of sample remaining in the sample vial should, in theory, load about 91% of the sample to the column.

After sample injection, peptides are separated by a solvent gradient optimized for chromatographic resolution. However, the gradient slope can significantly affect the detectability of peptides in LC-MS/MS. If a fast gradient is used, a peptide elutes quickly to form a fast rising peak in an ion chromatogram, resulting in intense signals in both MS and MS/MS spectra. But, in this case, only a few MS and MS/MS spectra can be acquired within the peak elution time. If a slow gradient is used, the same peptide would elute out more slowly to form a broader peak and the mass spectral signal of the peptide would be less intense. If a sufficient amount of sample is injected, the peptide signal intensity may be adequate to generate a database-searchable MS/MS spectrum. One major advantage of using a slow gradient for peptide elution is that a greater number of MS and MS/MS spectra can be acquired over this broad peak. For the analysis of a complex peptide sample, co-elution of different peptides cannot be avoided and one always tries to sequence as many co-eluting peptides as possible; a slow gradient provides this opportunity. However, if the amount of sample injected is small, the peptide signal may not be sufficiently intense to produce a database-

searchable MS/MS spectrum. Thus, the gradient slope needs to be optimized according to the sample amount injected to the LC-MS/MS instrument.

It was investigated how the gradient slope affects the number of peptides identified by LC-ESI MS/MS. It was found that the optimum gradient time increased as the number of cells in a sample increased. In addition, within a group of samples (e.g., the 500-cell samples), there was an optimal gradient time for detecting peptides. A gradient that was too long resulted in the identification of fewer peptides. Thus, the gradient time was adjusted according to the number of cells used for proteome analysis. Specifically, for the 500-cell samples, a 90-min gradient was used. The gradient time was increased to 150 min for the 1000-cell samples. The gradient time was 180 and 270 min for the 2500-cell and 5000-cell samples, respectively. Figure 2.3 shows the representative ion chromatograms generated from the trypsin digests of whole cell lysates of 500, 1000, 2500, and 5000 cells.

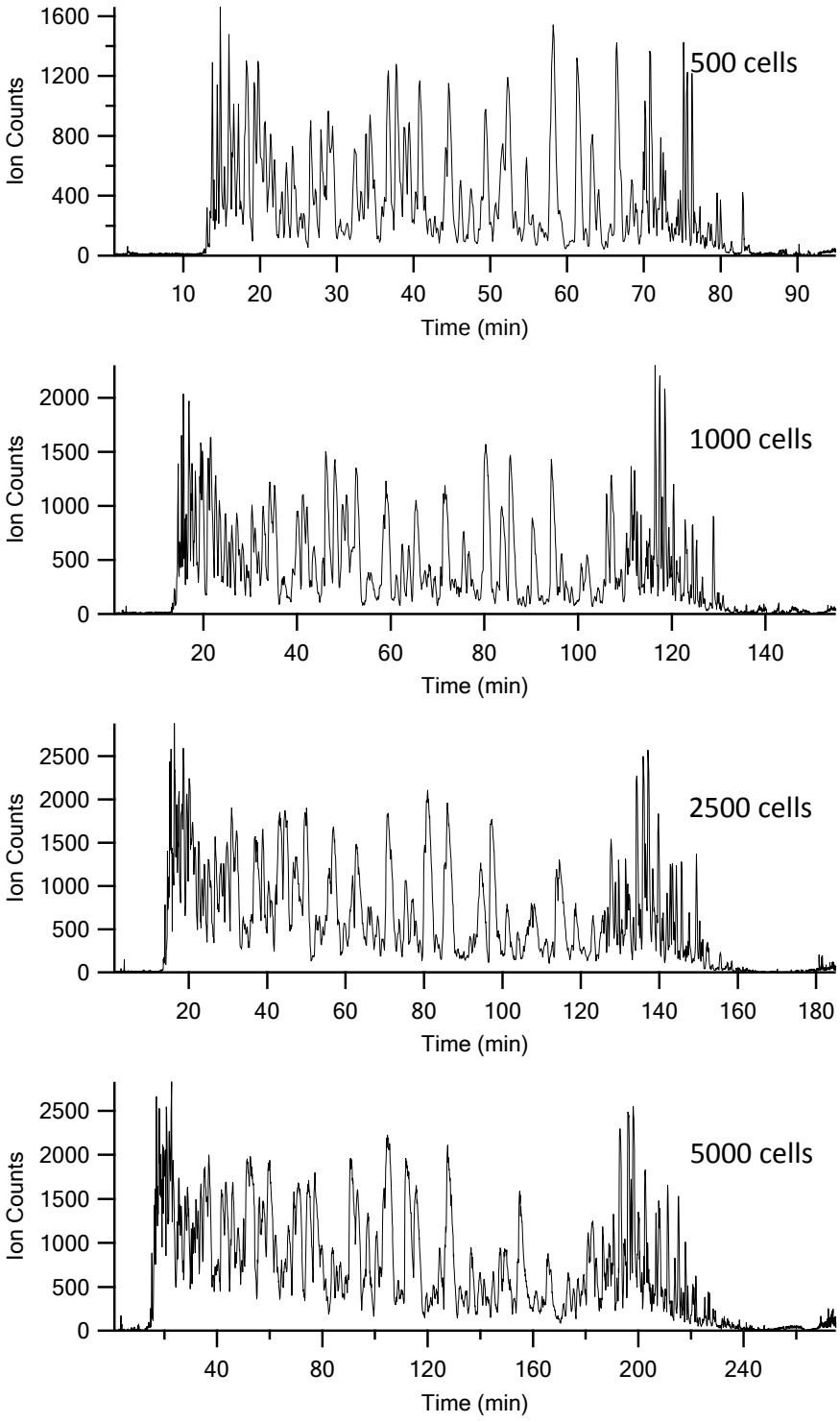

Figure 2.3 Base peak chromatograms from nano-LC QTOF MS/MS analysis of the trypsin digests from cell lysates of different numbers of cells.

Table 2.1 summarizes the peptide and protein identification results from the 500-, 1000-, 2500- and 5000-cell samples. In each group, three replicate experiments were carried out. The numbers of peptides and proteins identified from these samples are plotted in Figure 2.4. As Table 2.1 and Figure 2.4 show, both the numbers of peptides and proteins increase as the cell number increases and the number change is not in linear proportion to cell numbers. An average of 1891 ± 266 peptides or 619 ± 59 proteins (n=3) were identified from the 5000-cell sample. These numbers were compared favorably to 305, 211, 290, and 179 peptides or 113, 85, 133, and 83 proteins identified in four replicate runs of 5000-cell samples as reported by others.[22] The significant difference can be attributed to several factors including differences in sample handling, LC-ESI MS instrumentation and MS running conditions. In the case of 500-cell samples, 381 ± 11 peptides or 167 ± 21 proteins were identified using our method. Although the cell number decreases by 10-fold, compared to the 5000-cell sample, the number of peptides and proteins identified decreases by only about 5.0- and 3.7-fold, respectively. However, the peptide/protein ratio decreases from 3.05 for the 5000-cell sample to 2.14 for the 500-cell sample. These results indicate that an average of 167 proteins from 500 cells, 237 proteins from 1000 cells, 491 proteins from 2500 cells, and 619 proteins from 5000 cells can be identified. In all cases, the run-to-run reproducibility was good, indicating that the experimental protocol used in this study can be used to generate reproducible results from as few as 500 cells.

Table 2.1 Unique Proteins and Peptides Identified from Samples Containing Different Numbers of Cells.

Number of cells	Unique peptides	Unique proteins
500	369	168
	386	187
	389	145
1000	574	271
	485	226
	481	215
2500	1036	422
	1531	546
	1358	504
5000	1630	552
	2161	665
	1883	640

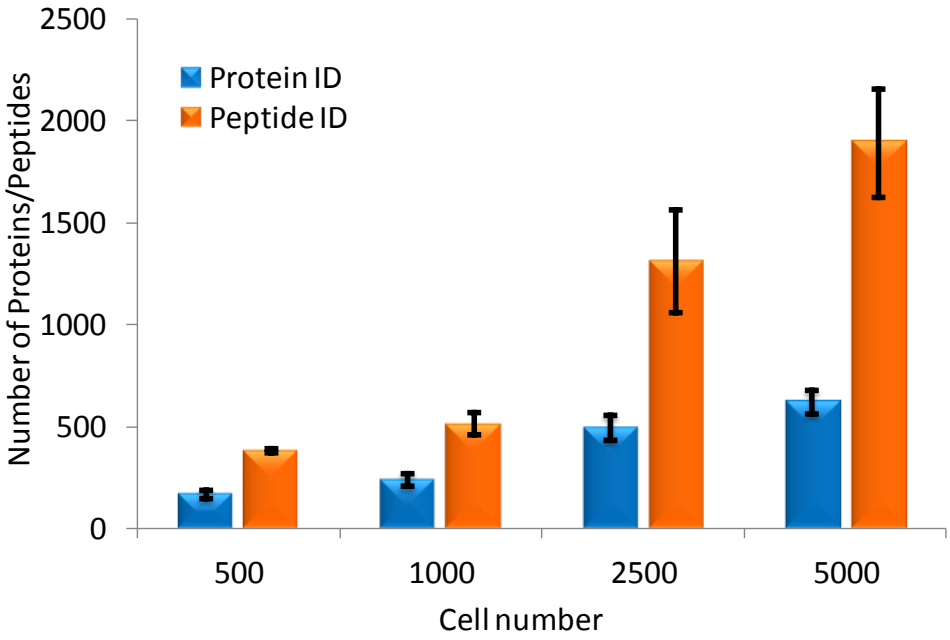

Figure 2.4 Protein and peptide identification results under optimized sample preparation and LC-MS/MS conditions.

While identification of one or more specific tumor biomarkers from small numbers of cells may prove to be useful for tumor diagnosis and progression monitoring, the clinical utility of proteome analysis from small numbers of cells may lie in the proteome profiling work. The ability to detect hundreds of proteins from as few as 500 cells using the current protocol opens the possibility of studying the proteome of a small number of cells such as CTCs isolated in blood of patients with cancer. Proteome profile may be used as a signature or fingerprint to identify a specific type of cancer cells in the human blood for detection, diagnosis and monitoring of cancer To mimic the scenario of analyzing CTCs in blood, a model system was used, where MCF-7 cells were spiked to fresh human blood, followed by isolation of these cells using density separation, antibody recognition and flow cytometry.

In this work, the erythrocytes were removed from peripheral blood leukocytes (PBL) and MCF-7 cells by using the Ficoll-Hypaque technique. This is a commonly used centrifugation technique for separating lymphocytes from other components in the blood according to their density differences. Studies have shown that the MCF-7 cells preferentially sediment with the PBL at the plasma and Ficoll interface based on their density differences.[23, 24] As a result, the spiked MCF-7 cells can be collected through the isolation of PBL from the interface after centrifugation. The buffy coat layer was then washed and re-suspended in a PBS buffer. The PBL cells are physically smaller than MCF-7 cells (averagely 8 μm comparing to 18 μm),[23] and their cell contents are also less complex than the cancer cells. These physical differences should be adequate for the flow cytometer

to differentiate the MCF-7 cells from PBL in cell sorting. However, to enhance the confidence of collecting the MCF-7 cells, FITC conjugated mouse anti-human HEA, an antibody specific to a human epithelial marker, was used.[25] In this case, the MCF-7 cells were stained with the fluorescent antibody whereas the PBL were not. Figure 2.5(A) shows the two-dimensional (2D) scatter plot of the flow cytometry analysis of the cell mixture. It clearly shows two populations. Population A represents the MCF-7 cells and population B represents PBL. To further guarantee only the cancer cells were collected, instead of the debris or aggregated cells, the gate for MCF-7 cell sorting was conservatively shrunk. Figure 2.5(B) presents the log scale fluorescence histogram of all the cells in the suspension. Given that only the MCF-7 cells are fluorescently labeled, population D should be the MCF-7 cells. A very small percentage of non-specific binding of the antibody to PBL was expected. However, with both gating strategies, shown in Figure 2.5(A) and (B), applied simultaneously during the flow cytometry analysis, the cancer cells were confidently sorted and collected.

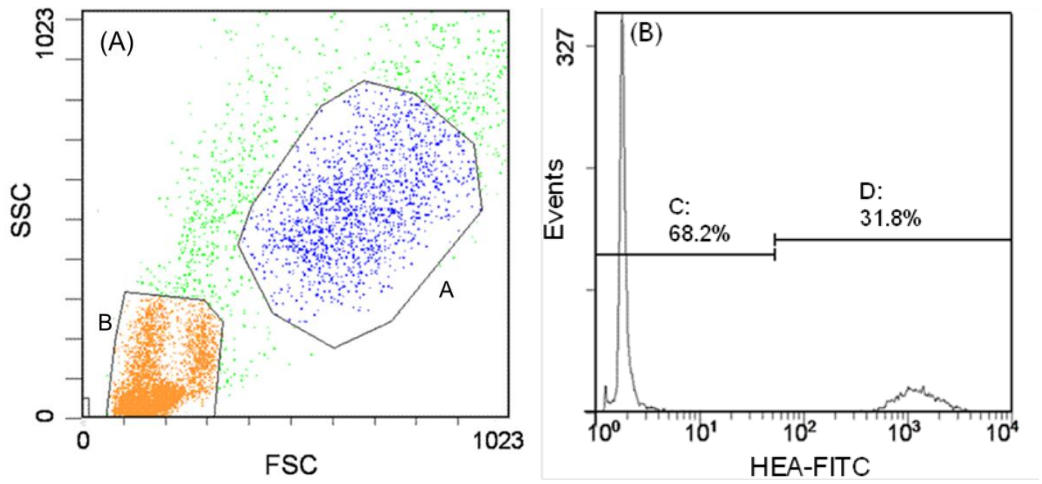

Figure 2.5 Flow cytometry results of the MCF-7 cells labeled with anti-HEA-FITC and the PBL cells. (A) 2D dot plot of MCF-7 and PBL mixtures where x-axis parameter (FSC) indicates the size of cells, and y axis (SSC) indicates granularity of cells. (B) the fluorescence response of both cells in the suspension where x-axis parameter (HEA-FITC) indicates intensity of FITC fluorescence in log scale.

The proteome profile of the isolated cells was then generated by the method described above and compared to those of the MCF-7 cell lines. The entire workflow for the isolation of the MCF-7 cells in blood is shown in Figure 2.2 and has been described in the Experimental section. Figure 2.6 shows the numbers of peptides and proteins identified from different numbers of cells isolated from the blood samples. The numbers are very similar to those obtained from the samples prepared directly from the cultured cells. Moreover, the proteome profiles are very similar, judging from the common proteins obtained from the two comparative samples (see Table 2.2). In Table 2.2, the results of intra- and inter-sample comparison (i.e., percentage of common proteins found in two samples) are listed. For example, in the case of 500 cells, three replicate experiments were carried for the 500-cell samples (Table 2.2 refers to them as A,

B, and C). Likewise, three replicate experiments were done for the 500-cell samples from blood spiked with MCF-7 cells (Table 2.2 refers to them as A', B' and C'). Within the dataset of A, B, and C, the average percent of common proteins found in two samples is 57% ± 10%. For the A', B' and C' samples, the average is 65% ± 11%. The average common protein percentage from the comparison of A vs. A', B vs. B', and C vs. C' is 60% ± 14%. The difference of these data is not significant. Thus, these proteome profiles are considered to be indistinguishable. This example illustrates that it is possible to generate a proteome profile from as few as 500 cells isolated from a blood sample and the proteome profile may be used for cell typing.

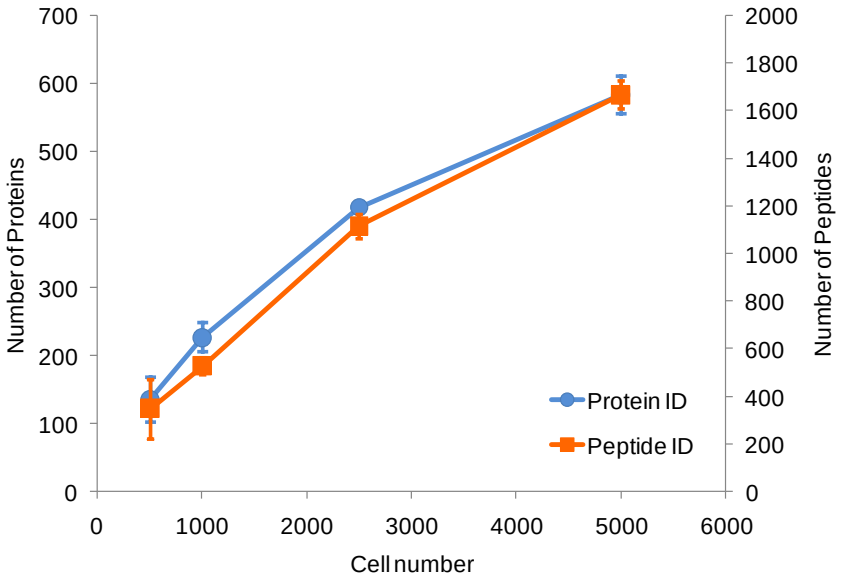

Figure 2.6 Protein and peptide identification results of MCF-7 cells isolated from a blood sample.

Table 2.2 Summary of Protein Identification Results from Different Runs.

	500 cells			1000 cells			2500 cells			5000 cells		
Sample	Overlap (%)*		Average	Overlap (%)*		Average	Overlap (%)*		Average	Overlap (%)*		Average
A&B	64	58		59	70		72	78		74	76	
B&C	57	73	57 ± 10	60	63	63 ± 6	64	77	72 ± 7	63	74	71 ± 6
C&A	41	48		56	70		60	78		64	77	
A'&B'	49	75		69	58		69	70		72	75	
B'&C'	73	67	65 ± 11	54	61	60 ± 5	71	70	69 ± 2	72	76	72 ± 3
C'&A'	51	72		59	56		67	66		67	73	
A&A'	48	47		51	66		57	74		66	72	
B&B'	49	82	60 ± 14	59	53	58 ± 5	62	76	68 ± 7	68	73	72 ± 4
C&C'	62	73		61	60		70	70		77	77	

*Percentage of common proteins found in two comparative runs. A, B, and C refer to the samples of three replicate experiments from the MCF-7 cells. A', B', C' refer to the samples of three replicate experiments from the cells isolated from blood spiked with the MCF-7 cells.

For the real world applications of this technique for generating proteome profiles of CTCs isolated from blood of patients with cancer, based on the study by Nagrath et al,[26] the low limit of 500-cell would render the current technique useful for 71 out of 115 patients (62%) with 10-mL blood collection per patient. This level of applicability should be useful in clinical sittings such as in longitudinal monitoring of cancer during treatment. However, if the technique could become more sensitive to generate similar level of proteome coverage (~167 proteins) using 200 cells, it might be applied to 98 out of 115 patients (85%). An attempt was made to analyze 250 cells using the protocol described above and only about 50 proteins were identified, a dramatic reduction in protein number considering the trend of gradual protein number decrease from 5000 cells to 500 cells. Careful inspection of the MS/MS spectra generated from 250-cell samples revealed that many of the spectra had some characteristic fragment ions

similar to those shown in their corresponding MS/MS spectra collected from a larger number of cells. Unfortunately, their Mascot ion scores were below the identity threshold. Thus, MS/MS database search did not match with any peptides. Our future work on technical development will focus on the sensitivity improvement of the current method.

2.4 Conclusions

A shotgun proteome analysis method has been developed for protein identification from thousands of cells. This method is based on the use of a detergent (NP-40) to disrupt the cells, followed by acetone precipitation. After washing the pellet with cold acetone to remove any residual detergent, the pellet was dissolved in NH_4HCO_3 and the solubilized proteins were subjected to trypsin digestion. By optimizing the sample volume, about 91% of the digest solution was injected into a capillary LC-ESI QTOF system for analysis. The resultant MS/MS spectra were searched against a proteome database for protein identification. In analyzing the MCF-7 cells, this method was demonstrated to be capable of identifying an average of 167 ± 21, 237 ± 30, 491 ± 63, and 619 ± 59 proteins from 500, 1000, 2500, and 5000 cells, respectively. This method was then applied to the analysis of proteome profiles of small numbers of cells isolated from a blood sample spiked with the MCF-7 cells. It was shown that the proteome profiles generated from the cells isolated in the blood sample were similar to those of the MCF-7 cells. We envisage that this method will be useful in proteome

profiling of small numbers of cells for disease diagnosis and prognosis. In addition, further optimization in the sample preparation process to reduce sample loss may result in identification of even more proteins from thousands of cells.

2.5 Literature Cited

(1) Espina, V.; Wulfkuhle, J.; Calvert, V.; VanMeter, A.; Zhou, W.; Coukos, G.; Geho, D.; Petricoin, E.; Liotta, L. *Nat. Protoc.* **2006**, *1*, 586-603.

(2) Hutter, G.; Sinha, P. *Proteomics* **2001**, *1*, 1233-1248.

(3) Ladanyi, A.; Sipos, F.; Szoke, D.; Galamb, O.; Molnar, B.; Tulassay, Z. *Cytom. Part A* **2006**, *69A*, 947-960.

(4) de Roos, B.; Duthie, S.; Polley, A.; Mulholland, F.; Bouwman, F.; Heim, C.; Rucklidge, G.; Johnson, I.; Mariman, E.; Daniel, H.; Elliott, R. *J. Proteome Res.* **2008**, *7*, 2280-2290.

(5) Schneider, T.; Moore, L.; Jing, Y.; Haam, S.; Williams, P.; Fleischman, A.; Roy, S.; Chalmers, J.; Zborowski, M. *J. Biochem. Bioph. Meth.* **2006**, *68*, 1-21.

(6) Swerts, K.; Ambros, P.; Brouzes, C.; Navarro, J.; Gross, N.; Rampling, D.; Schumacher-Kuckelkorn, R.; Sementa, A.; Ladenstein, R.; Beiske, K. *J. Histochem. Cytochem.* **2005**, *53*, 1433-1440.

(7) Allan, A.; Vantyghem, S.; Tuck, A.; Chambers, A.; Chin-Yee, I.; Keeney, M. *Cytom. Part A* **2005**, *65A*, 4-14.

(8) Chosy, E.; Nakamura, M.; Melnik, K.; Comella, K.; Lasky, L.; Zborowski, M.; Chalmers, J. *Biotechnol. Bioeng.* **2003**, *82*, 340-351.

(9) Utz, P. *Immunol. Rev.* **2005**, *204*, 264-282.

(10) Harwood, M.; Christians, E.; Fazal, M.; Dovichi, N. *J. Chromatogr. A* **2006**, *1130*, 190-194.

(11) Umar, A.; Luider, T.; Foekens, J.; Pasa-Tolic, L. *Proteomics* **2007**, *7*, 323-329.

(12) Sitek, B.; Sipos, B.; Schulenborg, T.; Marcus, K.; Schmiegel, W.; Hahn, S.; Kloppel, G.; Meyer, H.; Stuhler, K. *Mol. Cell. Proteomics* **2006**, *5*, S147-S147.

(13) Marko-Varga, G.; Berglund, M.; Malmstrom, J.; Lindberg, H.; Fehniger, T. *Electrophoresis* **2003**, *24*, 3800-3805.

(14) Seshi, B. *Proteomics* **2007**, *7*, 1984-1999.

(15) Wang, H.; Qian, W.; Mottaz, H.; Clauss, T.; Anderson, D.; Moore, R.; Camp, D.; Khan, A.; Sforza, D.; Pallavicini, M.; Smith, D. *J. Proteome Res.* **2005**, *4*, 2397-2403.

(16) Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* **2009**, *81*, 1049-1060.

(17) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(18) Wang, N.; Li, L. *Anal. Chem.* **2008**, *80*, 4696-4710.

(19) Fournier, M. L.; Gilmore, J. M.; Martin-Brown, S. A.; Washburn, M. P. *Chem. Rev.* **2007**, *107*, 3654-3686.

(20) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. *Nat. Biotechnol.* **2003**, *21*, 532-538.

(21) Speers, A. E.; Wu, C. C. *Chem. Rev.* **2007**, *107*, 3687-3714.

(22) Wang, H.; Qian, W.-J.; Mottaz, H. M.; Clauss, T. R. W.; Anderson, D. J.; Moore, R. J.; Camp, D. G., II; Khan, A. H.; Sforza, D. M.; Pallavicini, M.; Smith, D. J.; Smith, R. D. *J. Proteome Res.* **2005**, *4*, 2397-2403.

(23) Chosy, E. J.; Nakamura, M.; Melnik, K.; Comella, K.; Lasky, L. C.; Zborowski, M.; Chalmers, J. J. *Biotechnol. Bioeng.* **2003**, *82*, 340-351.

(24) Partridge, M.; Phillips, E.; Francis, R.; Li, S. R. *J Pathol* **1999**, *189*, 368-377.

(25) Hager, G.; Cacsire-Castillo, T. D.; Schiebel, I.; Rezniczek, G. A.; Watrowski, R.; Speiser, P.; Zeillinger, R. *Gynecol. Oncol.* **2005**, *98*, 211-216.

(26) Nagrath, S.; Sequist, L. V.; Maheswaran, S.; Bell, D. W.; Irimia, D.; Ulkus, L.; Smith, M. R.; Kwak, E. L.; Digumarthy, S.; Muzikansky, A.; Ryan, P.; Balis, U. J.; Tompkins, R. G.; Haber, D. A.; Toner, M. *Nature* **2007**, *450*, 1235-1239.

Chapter 3

Validation of Peptide MS/MS Spectra Using Metabolic Isotope Labeling for Spectral Searching-Based Shotgun Proteome Analysis*

3.1 Introduction

In mass spectrometry-based shotgun proteomics, the correlation between a tandem mass spectrometry (MS/MS) spectrum and a peptide sequence is a crucial step. Several sophisticated database search engines, such as Mascot[1], SEQUEST[2] and X!Tandem[3], have been developed to find the best match by comparing the experimental spectrum with the theoretical fragmentation patterns of individual peptide sequences derived from the protein sequences in a proteome database. The resultant matches are often assessed using statistical tools, either individually or globally, to arrive at a final list of peptide sequences that are deemed to be correct identifications at a defined confidence level.[4] While this strategy is widely used for proteome analysis, it has an inherent limitation: the intensity pattern of the fragment ion peaks in an MS/MS spectrum is difficult to predict and thus not fully utilized during the matching process. As an alternative, spectral library

*A version of this chapter has been published as Mingguo Xu and Liang Li, **2011**, "Validation of peptide MS/MS spectra using metabolic isotope labeling for spectral matching-based shotgun proteome analysis", *J. Proteome Res.* 10, 3632-3641.

searching shows great potential to address this problem. In this method, a spectral library is constructed by compiling the MS/MS spectra that have been identified and linked to specific peptide sequences. Unknown peptide identification is based on the comparison of its MS/MS spectrum with the library spectra. There are several reports of using this strategy for shotgun proteome analysis[5-14] with demonstrated advantages over the sequence database searching approach in terms of sensitivity, specificity and speed. If a reliable spectral library and a robust matching algorithm are at hand, it might become a better tool to analyze the proteome of a small number cancer cells (see Chapter 2). However, the success of this method is very much dependent on the construction of a reliable spectral library.

There are a few reported methods for compiling shotgun proteomics data to construct spectral libraries.[6, 7, 15] For example, Frewen *et al.*[7] used the SEQUEST search results of experimental MS/MS spectra with stringent scoring criteria to compile a list of high-score matches that were used to build the spectral library. For the redundant spectral identifications, spectral similarity was examined to select the most similar replicate spectrum as a library spectrum. Craig *et al.*[6] reported the use of X!Tandem to identify peptide sequences that were compiled into a publicly available database: GPMDB. In their study, for the redundant spectral identifications, the best match (i.e., the one with lowest expectation value) assigned by X!Tandem was chosen as the representative spectral identification in the library. In addition, to minimize the library space and speed up the spectral searching calculation, only the 20 most intense peaks were

retained in a library spectrum. While both methods provided a straightforward way to construct the spectral library, the spectra or fragmentation patterns of the matched peptides were not validated. Thus, some of the best matches may not be correct. And the most similar replicate spectrum may not truly represent the fragmentation pattern of a matched peptide sequence.

More recently, Lam *et al.*[12] reported a software tool for construction of customized spectral libraries. In their approach, instead of choosing one of the replicate spectral identifications as the representative spectrum in the library, a consensus spectrum was created by combining the similar spectra that have the same peptide identification. It was demonstrated that the consensus spectrum was a superior and more truthful representation of the fragmentation pattern of the peptide ions than the most similar replicate. It reduced the chances of false positive spectra being included in the library by discarding dissimilar peak patterns and collecting most of the common fragmentation information in the spectra. Moreover, in their workflow, PeptideProphet[16] was implemented to statistically validate the sequence-database search results before including them in the final spectral library. This provided an additional quality control to ensure only the most likely correct spectrum-to-sequence assignments were entered in the library.

Considering the rapid advances in developing algorithms for assessing the sequence-database search results, there is no doubt that even more powerful statistical tools will likely be developed in the near future for constructing spectral

libraries. However, statistical tools do not provide validation of the spectrum-to-sequence assignments. In this work, a strategy is reported that uses differential isotope labeling of proteins combined with trypsin digestion and two-dimensional liquid chromatography quadrupole time-of-flight mass spectrometry (2D-LC QTOF MS) to provide experimental evidence to validate the peptide-spectrum matches (PSMs) generated by sequence-database searching.

In this study, metabolic labeling was introduced to culture the cells in normal or ^{15}N-enriched media. This ^{15}N-labeling method has been applied to detect differentially expressed proteins in various proteomic systems, including cells from plants[17], cell lines[18], *C.elegans*[19], and even tissue of mammals[20]. The unlabeled and labeled peptides from the digests of the cellular protein extracts behave the same during the ionization and fragmentation process in a mass spectrometer. Previous study[21] demonstrated that by simply overlaying the unlabeled and ^{15}N-labeled MS/MS spectra of the same peptide sequence, experimental evidence can be provided to validate protein identification results generated by the sequence-database search method. A unique aspect of ^{15}N-labeling is that the fragment ions composed of different amino acid sequences have non-uniform mass shifts from their corresponding unlabeled fragment ions, facilitating the identification of correct assignment from other false matches. All amino acid residues contain nitrogen atoms, but not the same number. Thus, the fragment ions of two peptides of the same mass, but different sequences, will have different patterns of mass shifts in the overlaid spectral pairs. Background noise can be readily distinguished from the true peptide fragment ions as it would not

follow the same mass shift pattern (i.e., as a singlet). While this validation process is not the same as the direct comparison of an acquired spectrum with a spectrum generated from an authentic peptide standard, in the absence of a large number of peptide standards, the labeled peptides can be considered as the internal standards for spectral validation.

Herein a method is devised to use ^{15}N-labeling for validating the spectrum-to-sequence assignments generated from the sequence-database search using Mascot search engine to construct a more reliable MS/MS spectral library of a widely used model microorganism, *E. coli* K12. The experimental workflow and a data filtering strategy to construct the library are reported. The utility of this library for proteome analysis based on spectral search is demonstrated and the results are compared to the sequence-database search results.

3.2 Experimental

3.2.1 Chemicals and Reagents

Dithiothreitol (DTT), iodoacetamide (IAA), trifluoroacetic acid (TFA), urea and sodium bicarbonate were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, HPLC grade formic acid, LC-MS grade water, acetone, and acetonitrile (ACN) were obtained from Fisher Scientific Canada (Edmonton, Canada). The BCA assay kit was purchased from Pierce (Rockford, IL).

3.2.2 Sample Preparation

E. coli K12 (*E. coli*, ATCC 47076) was from the American Type Culture Collection. A single *E. coli* K12 colony was used to inoculate 30 mL of LB broth (BBL, Becton Dickinson). The culture was incubated overnight with shaking at 37 ˚C. About 11 mL of this cell culture was centrifuged at 300 *g* for 15 min and the cell pellets were added to 1 L of labeled or unlabeled growth medium in a 4 L baffled Erlenmeyer flask. Bio-Express Cell Growth Media was purchased from Cambridge Isotope Laboratories (Andover, MA) and the isotope purity of the unlabeled and ^{15}N-labeled media was 99%. Cells were harvested after 7 h of growth when the optical density at 600 nm was around 1.6 for both the unlabeled cells and ^{15}N-labeled cells, respectively. The cells were centrifuged at 400 *g* for 15 min and the cell pellets were resuspended in 45 mL of PBS buffer and passed twice through a minicell French press (Aminco, Rochester, NY) at 20,000 psi. About 5 mL of 10% Triton X-100 was added into the solution. After stirring for 20 min at 4 ˚C, cell lysates were frozen and stored at -20 ˚C. BCA assay on each aliquot of the cell lysate solution was performed to determine the protein concentration. Proteins in the cell lysates were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide (IAA). Acetone pre-cooled to -80 ˚C was added gradually to the cell lysates to a final concentration of 80% (v/v). The solution was then incubated at -20 ˚C overnight and centrifuged at 20,000 *g* for 15 min. The supernatant was decanted and the pellet was carefully washed once using cold acetone. The pellet was re-solubilized in 8 M urea. After 8-fold dilution to reduce the urea concentration to about 1 M, trypsin digestion was then carried out at protein to trypsin ratio of 50:1 at 37 ˚C for 48 h.

3.2.3 2D-LC MS/MS

Peptide mixtures were fractionated by strong-cation exchange (SCX) chromatography on an Agilent 1100 HPLC system (Palo Alto, CA) using a 2.1 × 150 mm PolySULFOETHYL A column with 5 μm diameter and 300 Å particle pore size (PolyLC, Columbia, MD). Desalting of each peptide fraction was performed on a 4.6 mm × 50 mm Polaris C18 A column with 3 μm particles and 300 Å pore size (Varian, CA). The eluted peptides were monitored and quantified using a UV detector operated at 214 nm.[22] The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, 1 μg of the digest was injected onto a 75 μm × 100 mm Atlantis dC18 column with 3 μm particles and 100 Å pore size (Waters, Milford, MA) via a Symmetry C18 trap column (180 μm × 20 mm). For the chromatographic separation, solvent A consisted of 0.1% formic acid in water and solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using a 120-min solvent gradient and introduced by electrospray into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A survey MS scan was acquired from m/z 350-1600 for 0.8 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 0.8 s each. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants (i.e., lock-masses), was infused at a flow rate of 300 nL/min, and a 1-s MS scan was acquired every 1 min throughout the run.

3.2.4 Mascot Search

Using Proteinlynx Global Server 2.3.0 (Waters) all raw LC-ESI data were lock-mass corrected, de-isotoped, and converted to peak list files with retention time information. All the peak list files were then submitted to Mascot search program (version 2.2.1). Database searching was restricted to *E. coli* K12 in the database. The search parameters for unlabeled samples were selected as follows: enzyme, trypsin; missed cleavages, 2; peptide tolerance, 30 ppm; MS/MS tolerance, 0.2 Da; fixed modification, carbamidomethyl (C); variable modifications, ammonia-loss (N-term C), N-Acetyl (protein), oxidation (M), carbamyl (N-term), carbamyl (K), pyro-Glu (N-term Q), and pyro-Glu (N-term E). The search parameters for the ^{15}N-labeled samples were the same as the unlabeled samples except the isotopic mass of nitrogen in all amino acids was set to be 15.0001. The search results, including original spectra, peptide sequences, retention time for each peptide identification, precursor m/z, ion score, Mascot threshold score for identity, rank in the result, calculated molecular mass of the peptide, corresponding protein names and access IDs were extracted to Excel files using an in-house program.

3.2.5 Metabolic Labeling Validation

Figure 3.1 shows the workflow for data processing to generate a list of validated spectra that were entered into the spectral library. Before going through the metabolic labeling validation step, intensity normalization was performed for all spectra by setting the intensity of the top 3 most intense peaks to 1 and rescaling the other peaks proportionally referenced to the intensity of the third most intense peak. The signal-to-noise (S/N) ratios of all the unlabeled spectra

were calculated by dividing the average intensity of the top ten most intense peaks by the median intensity of the spectrum.

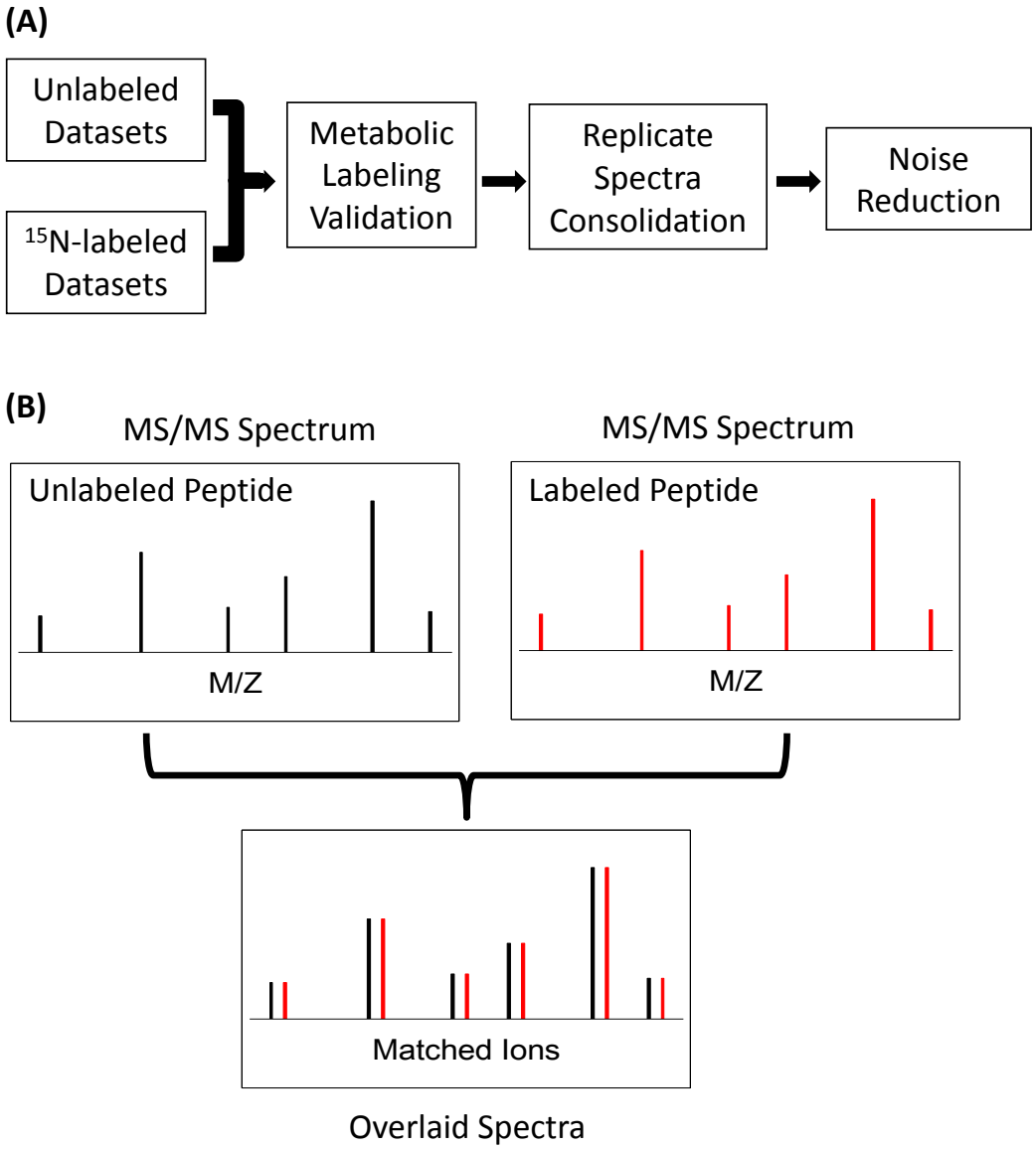

Figure 3.1 (A) Workflow of metabolic labeling validation of MS/MS spectra for constructing a spectral library. (B) Schematic of the process of overlaying an unlabeled peptide MS/MS spectrum with a labeled spectrum to determine the number of common fragment ions and similarity of the fragmentation patterns.

In the validation step the processed spectra from the unlabeled search results were overlaid with the ^{15}N-spectra of the same sequence matches. In the overlaid spectral pair, the identified fragment ions by Mascot were paired up based on the ion types (an example is shown in Figure 3.2). For each pair of the overlaid spectra, the number of common fragment ions was calculated and the similarity of the fragmentation patterns was calculated according to equation 1 (i.e., the spectral dot product of the fragment ion intensities). In equation (1),

$$\text{Dot product} = \frac{\sum(U_i \times L_i)}{\sqrt{\sum U_i^2 * \sum L_i^2}} \qquad (1)$$

L_i and U_i are the relative intensity of the same fragment ion, i, in the labeled spectrum and the unlabeled spectrum, respectively. For each spectral identification, only the top match with the highest number of common fragment ions and the highest similarity score was kept. Finally, two quality-control filters were applied to exclude the spectral identifications with less than 5 paired-up common fragment ions and with the similarity score of less than 0.96 (see Results and Discussion).

3.2.6 Replicate Spectra Consolidation

After validation, all the validated unlabeled spectra underwent a replicate-spectra consolidation process to construct a consensus spectrum for each peptide sequence assignment. Instead of simple averaging (arithmetic mean), the replicate spectral identifications (i.e., spectra with the same peptide assignment) were combined using weighted averaging (weighted mean). The weight for each spectrum was the calculated signal-to-noise (S/N) ratio itself. In this case, S/N

was calculated by dividing the average intensity of the top 10 most intense peaks in the spectrum by the median intensity of the spectrum. In this way, the better-quality replicates contributed more in the consensus spectrum.

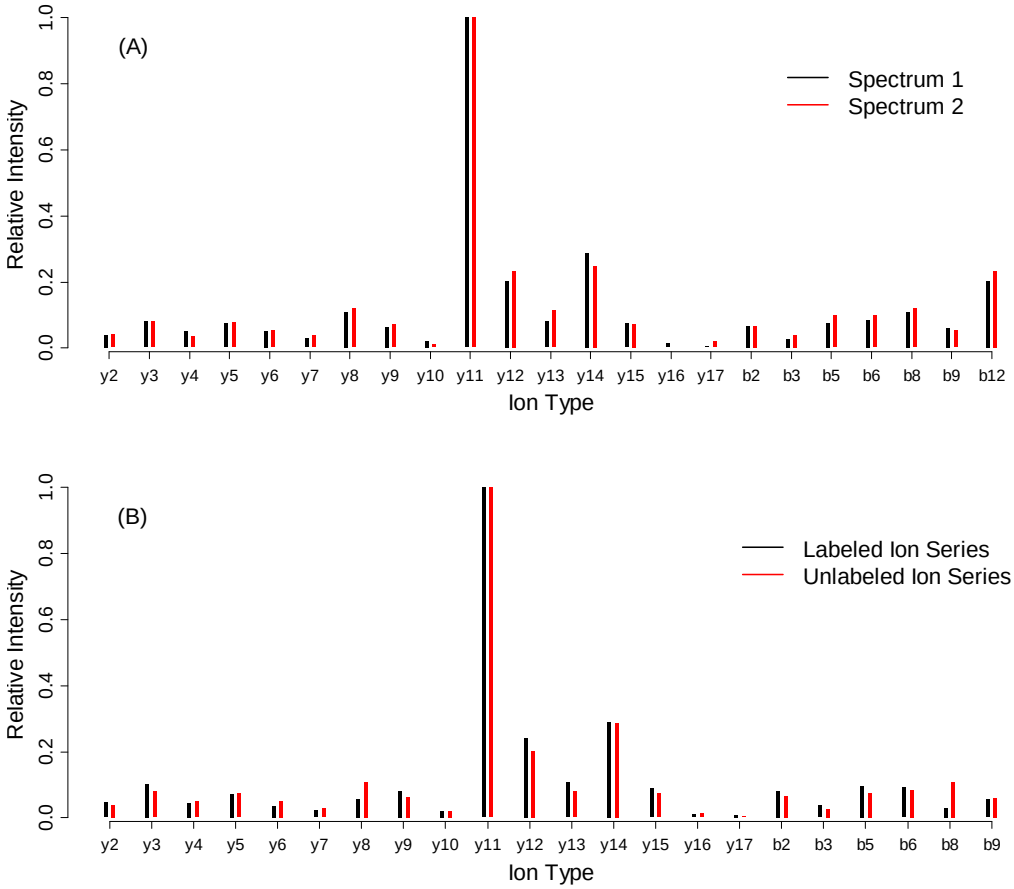

Figure 3.2 (A) shows the plots of relative intensity as a function of ion types from two unlabeled spectra (replicate) matched to the same peptide (TVINQVTYLPIASEVTDVNR). As expected, since these two spectra were good quality replicates, the fragmentation patterns of the two spectra look very similar. In fact, the spectral dot product calculated between the two spectra was 0.997. By comparison, (B) shows the plots of relative intensity as a function of ion types from a pair of unlabeled and labeled spectra. The fragmentation patterns of the common ion series are very similar with a spectral dot product value of 0.995. This representative example illustrates that the ^{15}N-labeling strategy would not affect the fragmentation pattern of peptide ions. Therefore, the similarity of fragmentation pattern between labeled and unlabeled can also provide experimental evidence to gauge the quality of the validation process.

3.2.7 Noise Reduction

All the validated fragment ions were kept for each spectral identification entered into the spectral library. In addition, noise reduction was performed by removing the peaks with m/z higher than the MH^+ peak (most likely to be mis-deisotoped peaks and contaminant peaks). Finally, all the invalidated peaks, including unidentifiable ions and invalidated identifiable ions, were sorted by their intensities. In order to simplify the library spectra, the maximum peak number per spectrum was set to be 100. Most invalidated peaks with low intensities were rejected.

3.2.8 Spectral Searching Algorithm

A spectral searching algorithm (SpecMatching) was developed to match the measured MS/MS spectra in the validated consensus spectral library. After being processed by Proteinlynx Global Server (Waters), peak lists with retention time were obtained from the raw data files. During the spectral searching process, intensity normalization was first performed for all the unknown spectra by setting the intensity of the top 3 most intense peaks to 1 and rescaling the other peaks proportionally. To reduce the variation in intensity measurement, square-root transformation of intensity was applied to all the fragment ions.[23] Next, the m/z value of the precursor ion was used to locate the library spectra with similar precursor m/z values for spectral similarity comparison. The peak list of a spectrum was divided into a consecutive sequence of 1 Th-wide bins on the *m/z* axis ranging from 50 Th to MH^+, and slightly shifted peaks were binned into the adjacent bin. The vector of the spectrum was calculated by summing the weighted

intensities of the peaks in each bin. Invalidated peaks and validated peaks were assigned different weighing factors by multiplying their intensities by 1 and 2, respectively. In the algorithm, a modified spectral dot product was used to measure the spectral similarity as the equation (2) shown below:

$$DP = \frac{\sum(M_i \times L_i)}{\sqrt{\sum M_i^2 * \sum L_i^2}} \qquad (2)$$

where L_i and M_i are the weighted intensity of the i^{th} bin of the library spectrum and that of the matching bin (of the same *m*/*z* value) of the measured spectrum, respectively.

3.2.9 Statistical Analysis

In Mascot search results, target-decoy search strategy[24] was applied by searching the MS/MS spectra against the forward and reverse *E. coli* K12 proteome sequences to calculate the false discovery rate (FDR). For plotting receiver-operating characteristic (ROC) curve, a Perl script from Matrix Science website (http://www.matrixscience.com/help/decoy_help.html) was used to tabulate the FDR data.

In the spectral searching results, since the negative and positive distributions were found to be properly fitted by a bi-normal distribution (see Results and Discussion), the number of true positive, true negative, false positive and false negative identifications could be estimated at a chosen score threshold. Using the equation (3) below, the FDR of the search results was calculated:

$$\text{FDR} = \frac{\text{False Positives}}{(\text{False Positives} + \text{True Positives})} \quad (3)$$

3.2.10 Software Development

All in-house software was written in Perl 5.12 (http://www.perl.org). Statistical analysis was performed using both Microsoft Excel as well as R scripting. Charts and graphs were generated using R's plotting packages (http://www.r-project.org/). Software was run on standard desktop and laptop computers running Windows 7 (Home Edition).

3.3 Results and Discussion

The main purpose of this study was to develop an experimental means to validate peptide identifications obtained from conventional sequence-database searching using MS/MS spectra in order to construct a more reliable peptide spectral library. The overall workflow is shown in Figure 3.1A. The peptide identifications were first generated using Mascot searches from two sets of samples, i.e., the proteome digests of proteins extracted from cells cultured in ^{14}N- or ^{15}N-metabolic media. A set of filtering criteria were applied to examine the two datasets, producing a list of spectra considered to be validated. The replicate spectra of the same sequence were consolidated to build a consensus spectrum. Both the individual spectrum from a single spectrum identification and the consensus spectrum from the redundant identifications were further processed to

exclude noise peaks. The final list of the validated unlabeled spectra was compiled to form the spectral library. The ^{15}N-metabolic labeling method was chosen for validation of the Mascot search results. As shown in Figure 3.1B, by overlaying the two MS/MS spectra obtained from the unlabeled peptide and its corresponding ^{15}N-labeled peptide, the number of common fragment ions (i.e., y- and b-ion series) and the intensity similarity between the common ions can be readily determined.

3.3.1 Mascot Result Analysis

Table 3.1 summarizes the results obtained from the 2D-LC separation and QTOF MS/MS analysis of the *E. coli* K12 samples. In total, 257,907 and 245,156 spectra were collected from the unlabeled and ^{15}N-labeled samples, respectively and 181,533 and 192,649 spectra were found to contain peptide identifications in the unlabeled and ^{15}N-labeled dataset. By applying a confidence level of 99% to both datasets, 37,699 and 33,095 peptides were matched, corresponding to 10,414 and 9,340 unique peptide sequences for unlabeled and labeled datasets, respectively. The estimated false-discovery rates (FDRs) were 1.12% for the unlabeled dataset and 2.21% for the labeled dataset.

Table 3.1 Summary of the Results Obtained from the Unlabeled and ^{15}N-labeled *E. coli* K12 Whole Cell Lysate Digests.

	Unlabeled dataset	^{15}N-labeled dataset
Total spectra	257,907	245,156
Spectra with peptide matches	181,533	192,649
All peptides (99% CL)*	37,699	33,095
Unique peptides (99% CL)*	10,414	9,340
Peptides matches (score>=13)	69,696	93,971
False-discovery rate	1.12%	2.21%

*confidence level (CL).

Figure 3.3 shows the score distributions of all the possible peptide matches in the two datasets including those with one peptide-spectrum match (PSM). As shown in Figure 3.3A, the Mascot scores for the unlabeled dataset range from 0 to 225, with a majority of matches having scores of lower than 10. Similar distribution can be seen for the ^{15}N-labeled dataset (Figure 3.3B); the Mascot scores range from 0 to 246, with a majority of scores of lower than 10. If a global FDR of 1% was applied to construct a spectral library, which corresponded to a Mascot identity threshold of 32 in the unlabeled dataset, 81% of the peptide matches would be discarded. While such a high threshold filter can reject most incorrect PSMs, it can also potentially over-exclude many correct peptides, resulting in reduced sensitivity. Even with this stringent filtering, there is no guarantee that the retained spectra are assigned correctly to the peptide sequences. Global FDR is only a statistical estimation of the matching quality for the whole dataset, not an assessment of each PSM. In this approach, an experimental method was used to examine all the PSMs to determine the correct identifications.

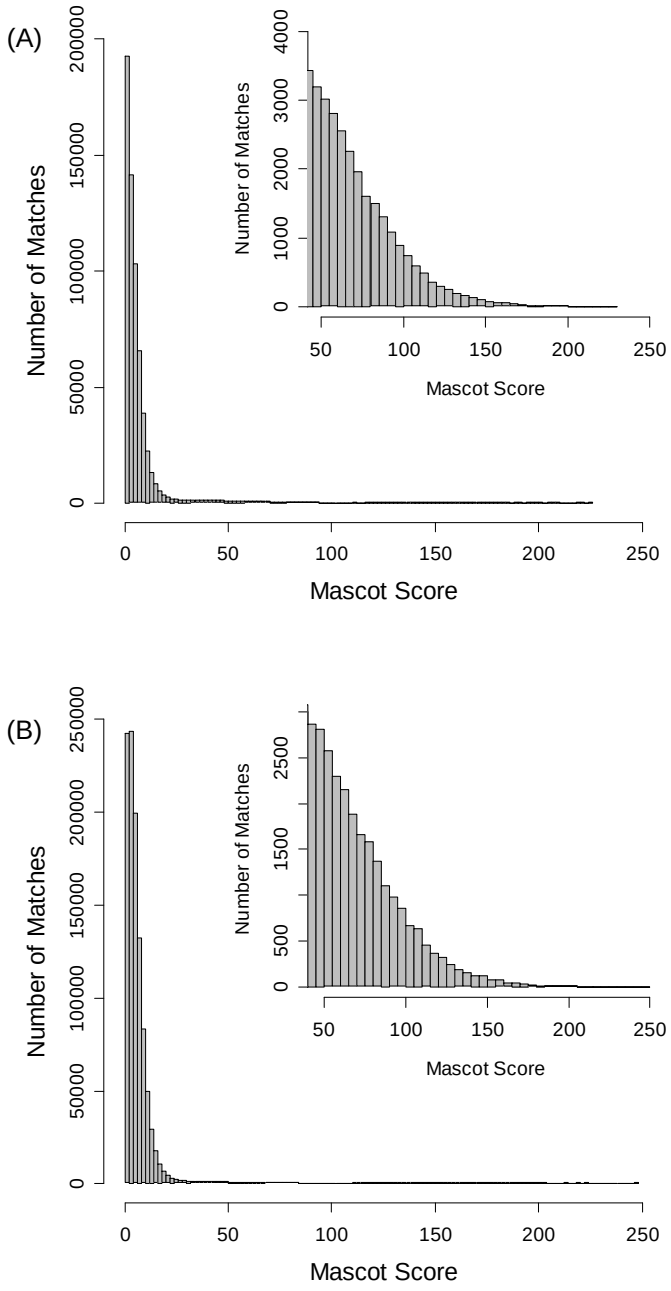

Figure 3.3 Mascot score distributions of all the possible peptide matches in the datasets of (A) unlabeled sample and (B) labeled sample, including those with one peptide spectral match (PSM). The insets are the expanded regions showing the score distributions.

3.3.2 Data Filtering for Validation

As indicated in Figure 3.1A, all the peptide identifications from the unlabeled and labeled datasets had undergone metabolic labeling validation. Briefly, if a spectrum from the unlabeled dataset had the same peptide identification and charge state as the one from the ^{15}N-labeled dataset, they would be overlaid as illustrated in Figure 3.1B. The number of common fragment ions (i.e., b- and y- ions and their corresponding neutral loss ions) in the overlaid spectra was determined and the intensity similarity between these common ions was calculated. In total, 271,253 comparisons were done, including comparisons between redundant peptide identifications. Each comparison contained one overlaid spectral pair, such as the one depicted in Figure 3.1B. Theoretically, in an overlaid spectral pair the labeled and unlabeled spectra of the same peptide sequence would generate the same fragment ions because the isotope labeling does not alter the chemical property. Therefore, it is reasonable to state that, the higher the number of common ions (only b- and y-ion series are considered) shared by the labeled and unlabeled spectra, the higher the confidence of these spectral identifications. Thus, the first filter used to determine whether a matched spectrum is a correct one or not is to count the number of common ions between the two spectra in the overlaid spectral pair.

Figure 3.4A shows a plot of the number of comparisons as a function of the number of common fragment ions found in the overlaid spectral pair. The number of common ions ranges from 0 to 35 and most of the overlapped identifications share 5 to 20 common fragment ions. Some of them have few

common ions (<5). There are two possible scenarios to have few common ions. The first possibility is that at least one of the identifications in a pair of spectra is false. The identified peaks in the spectrum are not from the real fragment ions, but noise peaks or fragment ions of other peptide ions with similar precursor ion masses. One unique feature of using differential ^{14}N- and ^{15}N-labeling is that the m/z shifts of the fragment ions between the two spectra are not uniform; the shift is dependent on the composition of the amino acids or nitrogen number in the fragment ion.[21, 25, 26] Thus, coincidental m/z match(es) between the unlabeled and labeled peaks may arise from only a few ions. The second possibility is that both identifications are correct, but due to some reasons (e.g., difficult to dissociate, low peptide concentration, etc), one of the spectra or both spectra do not have many y- and b-ions to begin with. Since these low quality spectra cannot truthfully represent the fragmentation pattern of their corresponding peptides, they should not be kept in a validated spectral library for high confidence peptide identification. The confidence level clearly increases as the number of the common y- and b-ions in a pair of spectra increases. In order to set up the minimum number of common ions used to filter the Mascot search results, the relation between the number of matched ions and the Mascot score was investigated.

Figure 3.4B shows the plot of the number of common fragment ions found in individual overlaid spectral pairs as a function of the average Mascot score of the corresponding peptide identifications ranging from 0 to 80. As expected, there is a strong positive correlation between the number of common ions and the

Mascot score. The curve can be fitted by a linear equation ($y = 0.1019x + 3.4467$) with the R^2 value equals to 0.9899. The Mascot identity threshold score is 13 at a significance level of 0.05 (i.e., p=0.05) when searching the unlabelled dataset against the *E. coli* proteome sequence database. As Figure 3.4B shows, when the Mascot score is 13, there are, on average, 5 matched fragment ions. Therefore a cut-off threshold of 5 common ions was adopted in the validation process. In other words, if the overlaid spectral pair contains less than 5 common ions, the peptide identification and the spectra will not be entered into the final validated list. It should be noted that the use of 5 common ions as the cut-off does eliminate some high-score matches and some of them may well be correct matches, as can be seen from Figure 3.4B. But this compromise was taken, considering that an accepted library spectrum with a smaller number of fragment ions may not be very useful for a spectral match based on spectral similarity calculation (see below); decreasing the number of fragment ions used for matching increases the chance of a false match with noise or impurity peaks.

As Figure 3.4B shows, the use of 5 common ions as the cut-off also includes some low-score ions. These matches as well as other high-score matches were examined by applying another filter which is based on the intensity similarity among the common fragment ions in the overlaid spectral pair. Since both unlabeled and ^{15}N-labeled datasets were collected by using the same mass spectrometer with the same instrument settings, the unlabeled peptides should have nominally the same fragmentation behavior as their ^{15}N-labeled counterparts, i.e., they should have the same fragment ions and similar intensity distribution

among these fragment ions. In this work, the similarity of fragmentation patterns is measured by calculating the spectral dot product of the common ions by equation (1) (see Section 3.2.5). Since the intensities of the fragment ions are all positive, one would expect the calculated scores to range from 0 to 1. A similarity score equal to 0 means that the two fragmentation patterns are unrelated, while a score of 1 means that the two fragmentation patterns are identical. An example of spectral similarity comparison is shown in Figure 3.5.

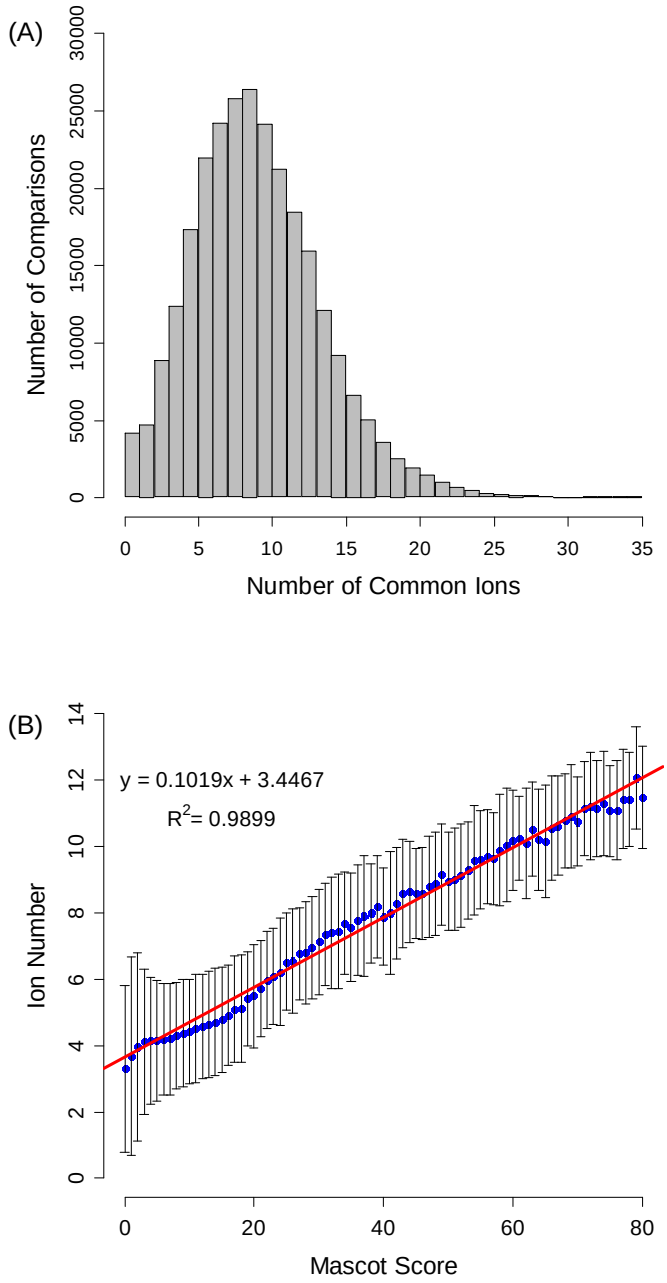

Figure 3.4 (A) Number of comparisons as a function of the number of common fragment ions found in the overlaid spectral pair. (B) Number of common fragment ions found in individual overlaid spectral pairs as a function of the average Mascot score of the corresponding peptide identifications.

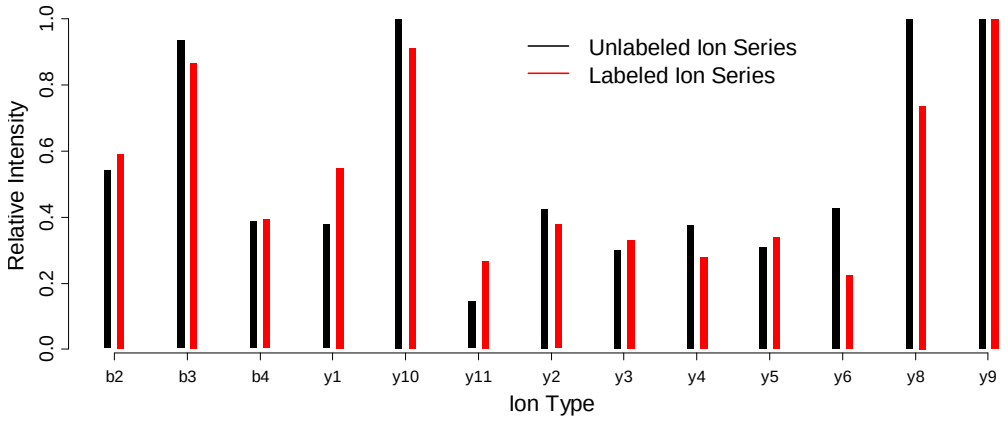

Figure 3.5 Relative intensity as a function of ion types from a pair of unlabeled and labeled spectra for the same peptide sequence ADDYTGPATDLLLK. Similarity score = 0.99.

This filter was applied to all overlaid spectral pairs and the resultant distribution of the similarity scores is shown in Figure 3.6A. As Figure 3.6A illustrates, a high similarity score (larger than 0.95) is obtained in most cases, which means that the intensity pattern between the common ions in the unlabeled and ^{15}N-labeled spectra is quite similar. They are from the correctly annotated ions, not random matches to noises or other ions. However, there are still some cases with relatively low scores which are more likely from random matches. To determine the similarity score threshold to be used as the second filter to exclude possible random matches, a control experiment was conducted to analyze the unlabeled sample in replicate and then the similarity of fragmentation patterns between the replicate identifications was examined using equation (1). Figure 3.6B shows the distribution plot which is very similar to Figure 3.6A. The distribution shown in Figure 3.6B reveals that 95% of the cases have a similarity score of 0.96 or higher. Thus, a similarity cut-off score of 0.96 was used as the

second filter to reject the potentially false peptide matches during comparisons of unlabeled and labeled spectral pairs.

In our work, all possible identifications were included in the validation process, regardless of their ranks in each spectral identification list in the Mascot search results. Thus, it was necessary to screen out some unreliable matches to make sure that only one peptide sequence was assigned to one spectrum. To do this, the number of common ions and the similarity score was used to make the judgment: only the match that had the highest common ion number and similarity score would be kept.

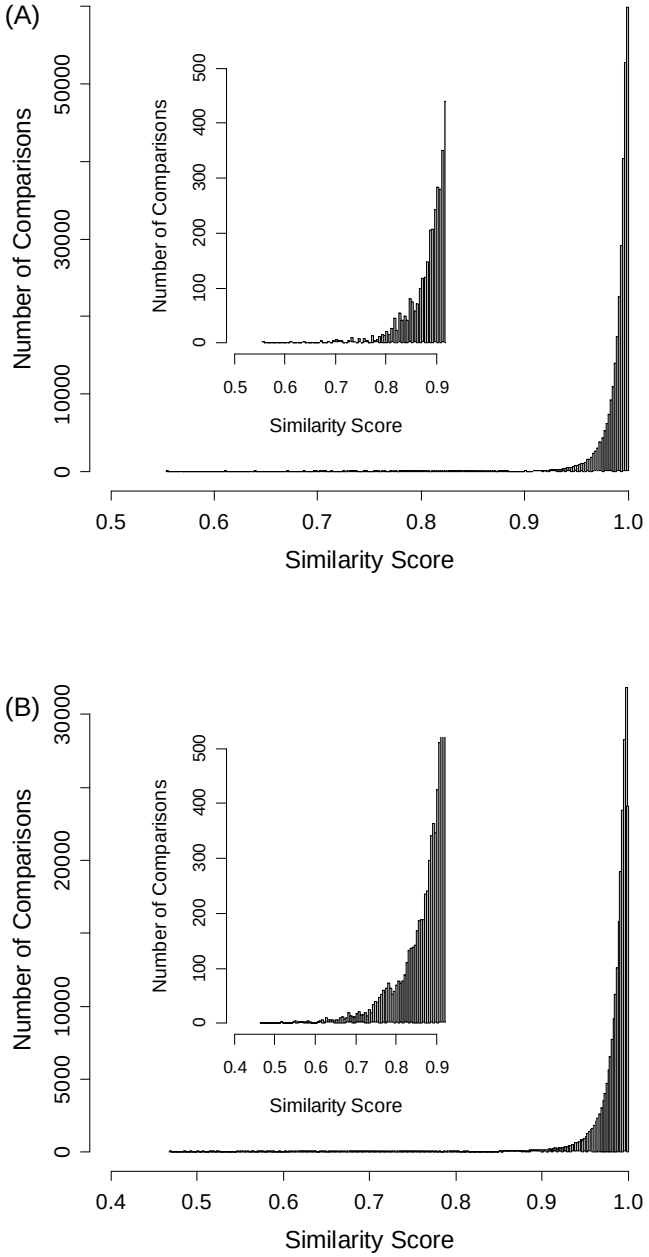

Figure 3.6 Number of comparisons as a function of similarity scores from (A) the comparison of unlabeled and labeled matches and (B) the comparison of the unlabeled matches from replicate runs.

In some cases, the 2nd or higher rank of matches in a spectral identification was found to be the correct one. For instance, one unlabeled

spectrum was found to match with the peptide sequences of ALEEAGAEVEVK or VPPGVDEAAYVK in the Mascot search result (see Table 3.2). After determining the numbers of common ions and the similarity scores with the labeled spectrum, the second-ranked peptide match had a higher number of common fragment ions (12 vs. 10) and higher similarity score (0.9875 vs. 0.7772) than the top-ranked counterpart. In addition, the top-ranked identification had a precursor mass error of 0.016 Da, while the second-ranked identification had an error of 0.0007 Da. Taken together the second-ranked match was deemed to be more reliable than the top-ranked match and, therefore, should be kept in the validated peptide identification list. It should be noted that the majority of kept peptide identifications (99%) were the top-ranked peptide assignments. Nevertheless this example shows that the top-ranked peptide with a high Mascot score was not always the correct identification.

Table 3.2 Summary of the Peptide Matches Obtained from Mascot Search of A MS/MS Spectrum and the Results Generated from the Validation Process.

Sequence	Rank	Mascot Score	Number of Common Ions	Similarity Score	Precursor Mass Error
ALEEAGAEVEVK	1	43.76	10	0.7772	0.016 Da
VPPGVDEAAYVK	2	33.63	12	0.9875	0.0007Da

3.3.3 Validated Spectral Library

After applying two validation filters and rank comparison, 54,447 overlaid spectral pairs were removed. In order to construct a consensus spectrum for each peptide sequence assignment, the remaining 216,806 overlaid spectral pairs underwent replicate-spectra consolidation and noise reduction. These two steps are very similar to the way proposed by Lam *et al.* for constructing SpectraST.[8, 12] However, instead of a 5 fold boost of the intensities of the annotated peaks in the consensus spectra, the intensity distribution of all the fragment ions was not artificially altered. It is believed that in this way the fragmentation pattern of a peptide ion has been more truthfully preserved, especially for the peptide ions that fragment irregularly (e.g., when the dominant ions are not the usual b- or y-ions). In our work, noise reduction was performed by removing the peaks with m/z of higher than that of the MH^+ peak. The high mass peaks are very likely to be mis-deisotoped peaks and contaminant ion peaks. Finally, all the invalidated peaks, including unidentifiable ions and invalidated identifiable ions, were sorted by their intensities. In order to simplify the consensus spectra, the maximum number of peaks per spectrum was set to be 100. Most invalidated peaks with low intensities were rejected. Since these peaks contribute little when measuring similarity between the measured spectrum and the consensus library spectrum in the spectral searching algorithm, it is reasonable to remove them.

After consolidation and noise reduction, an MS/MS spectral library of tryptic peptides was able to be constructed for *E. coli* K12. It constituted 9,302 unique PSMs (unique sequence and charge state) and 7,763 unique peptide sequences. These spectra were compiled and re-searched using Mascot. At the 99%

confidence level, the estimated FDR for this validated dataset was found to be 0.15%. Compared with the original result under the same condition (1.12% FDR), the FDR decreased dramatically. To keep the same global FDR for the experimental dataset without going through the metabolic labeling validation, the Mascot score threshold would need to increase to 43. With such a high threshold, only 6,722 unique peptide sequences would remain, i.e., 1,041 fewer than the unique peptide number obtained after undergoing the validation process. Moreover, by only applying the global FDR filter, the remaining PSMs would have no experimental support to validate their identifications.

The peptide matches from the Mascot search was compared with the entries in the validated spectral library. There were 5,669 common peptides found in the two datasets with 1,053 peptides found only in the Mascot results and 2,094 in the spectral library. The missing 2,094 peptides in the Mascot results were simply due to the false exclusion of the true positive identifications by setting a high Mascot score threshold. For instance, one spectrum was identified as ADDYTGPATDLLLK with a score equal to 31 by Mascot. It was rejected by the Mascot identity threshold filter. However, by going through our validation process, most of its assigned fragment ions in this spectrum, including 13 very intense y- and b-ions were able to be validated (see Figure 3.5). As shown in Figure 3,5, the fragmentation patterns between the unlabeled and ^{15}N-labeled spectra are quite similar, indicated by the similarity score of 0.99. This example demonstrates that by using the validation process some false negative

identifications could be retrieved from the discarded identifications for entering into the spectral library.

There are 1,053 peptide matches with high Mascot scores (higher than 43) that are not in the spectral library. Among them, 911 (87%) matches had only the unlabeled spectrum, missing the ^{15}N-labeled counterpart. Therefore they could not go through the validation process. To increase the number of matched pairs, which should result in the increase of validated spectra, it is necessary to develop an optimal precursor ion inclusion strategy, similar to the precursor ion exclusion strategy[27], where the precursor ion masses of the identified peptides from running the unlabeled sample will be used to direct the spectral collection of the same peptides in the labeled sample.

For the rest of the 142 matches, there were several reasons that they failed the validation process. First of all, the spectral quality of the unlabeled and ^{15}N-labeled spectra was quite different, with the labeled one having significantly poorer quality. Three examples are shown in Table 3.3. For instance, the peptide, QVEALVEASKEEVK, was identified in the unlabeled dataset with a Mascot score of 83.05, while the Mascot score of its ^{15}N-labeled counterpart was only 15.72. During the validation process, only 3 common y- and b-ions were found and their intensity patterns were very different, having a similarity score of 0.7709. Thus this peptide would not be considered as a validated sequence despite its high Mascot score in one of the spectral pair. It was found that 109 out of the 142 matches had low score counterparts and did not pass the validation process. Again, the use of a precursor ion inclusion strategy during the spectral collection of the

^{15}N-labeled sample may overcome this problem, as more time would be spent on generating the MS/MS spectra of the same peptides found in the unlabeled sample, thereby increasing the ^{15}N-labeled spectral quality.

Table 3.3 Examples of High-score Matches from the Unlabeled Peptides with Low-score Matches from the Labeled Peptides.

Sequence	Mascot Ion Score (Unlabeled)	Mascot Ion Score (Labeled)	Common Ions	Similarity Score
EAIHMYGPDYGFDTTINK	144.14	14.32	1	N/A
QVEALVEASKEEVK	83.05	15.72	3	0.7709
VAFTALVEK	79.58	19.2	6	0.8127

For the remaining cases, even though the unlabeled and ^{15}N-labeled pairs had similar Mascot scores, they still failed the validation process. They either did not share enough common y- and b-ions (12 matches out of the 142 matches) or their fragmentation patterns differed too much (21 matches).

From the above discussion, it can be concluded that the spectral library created using the proposed validation process can have more spectra entries than that created by the Mascot search. The missing spectra are mainly due to the absence of counterparts in the individual spectral pairs. Future work in reducing the number of such singlets, such as the use of precursor ion inclusion strategy, should overcome this problem.

3.3.4 Spectral Searching for Peptide Identification

To utilize the validated spectral library for shotgun proteome analysis, a spectral searching algorithm called SpecMatching was developed. As indicated in the Experimental Section, an optimized spectral dot-product equation (equation 2) was implemented to measure the similarity between the measured spectrum and the library spectrum. To demonstrate the performance of this method, a mixture of tryptic peptides from *E. coli* whole cell lysates was prepared and analyzed by LC-ESI-QTOF MS. All the generated spectra were searched using both Mascot and SpecMatching.

In SpecMatching, the results table reports matches with similarity scores ranging from 0 to 1. All the calculated scores were collected, sorted from small to large, and plotted as the score distribution shown in Figure 3.7A. This figure shows clearly that there are two distinct score populations slightly overlapping with each other at the scores between 0.5 and 0.7. The well-separated two populations indicate great discriminatory power of SpecMatching to distinguish the correct and incorrect peptide matches. Therefore, a similar idea was applied to that used in PeptideProphet[28] to fit this curve with two distribution functions. It was found that normal distribution offers a close approximation to both score distributions. After fitting these two distributions by using the maximum likelihood estimation package in R, a probability-probability plot or p-p plot was generated and is presented in Figure 3.7B to show the goodness of fit. In Figure 3.7B, the observed probability in the histogram is plotted against the computed probability and shown as a dashed line. As shown in Figure 3.7B, the solid line

almost completely overlaps with the 45 º solid line. This indicates that the computed probabilities are an accurate reflection of the observed probabilities.

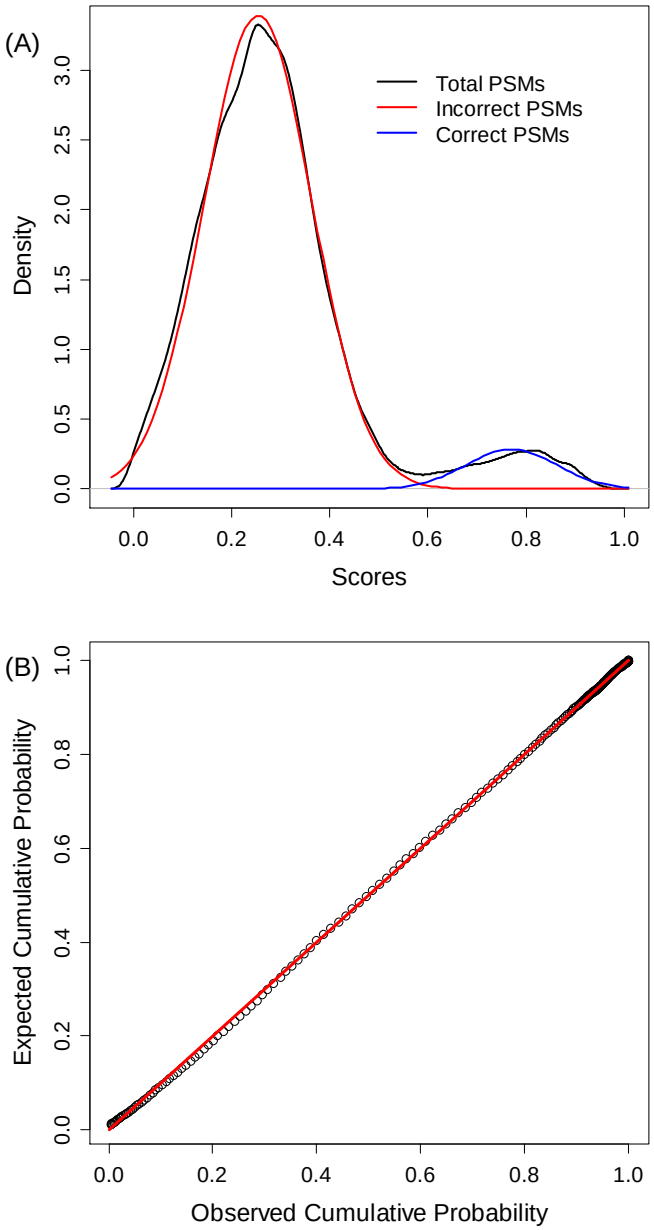

Figure 3.7 (A) Similarity score distribution of the matched peptides from a test sample by using SpecMatching against the validated E. coli spectral library. (B) Probability-probability plot showing the goodness of fit using two normal functions combined to represent the overall score distribution shown in (A).

Since the positive and negative distribution can be simulated by two normal distributions, one can readily calculate the different contributions at the overlapped portion of the histogram and estimate both the global and local FDR at a given score cut-off (see Figure 3.8A).

The same dataset was searched using Mascot in order to compare the result with the SpecMatching result. The receiver-operating characteristics curves (ROC curves) were plotted for both search methods (Figure 3.8B). The results clearly show that SpecMatching is able to identify more peptides at all desired false-discovery rates than Mascot, which indicates superior sensitivity of SpecMatching to Mascot. Thus, SpecMatching has better discriminatory power to differentiate correct and incorrect peptide matches than Mascot. This can be attributed to the fact that the spectral searching algorithm uses the intensity pattern of all fragment ions more truthfully than the sequence searching algorithm does.

A direct comparison was also carried out for the peptide matches (only the top-ranked matches) from both search results. In this case, a conservative score threshold was chosen for both search results. For the Mascot result, 0.01 of significance threshold was chosen, meaning that 1% of the peptide identifications in the result might be false. The global FDR for this Mascot result was estimated to be 0.4%. For the SpecMatching result, the threshold was gauged by either global or local FDR. For a fair comparison, 0.4% global FDR was chosen for the SpecMatching result, meaning that in the result 0.4% of peptide matches might be false spectral identifications. As Figure 3.8B shows, the number of spectral

identifications above the specified threshold in SpecMatching was 1,205 (979 unique peptide sequences). Among those peptide matches, 711 peptides can be found in both results, which constitute around 88% of the Mascot result and 73% of the SpecMatching result.

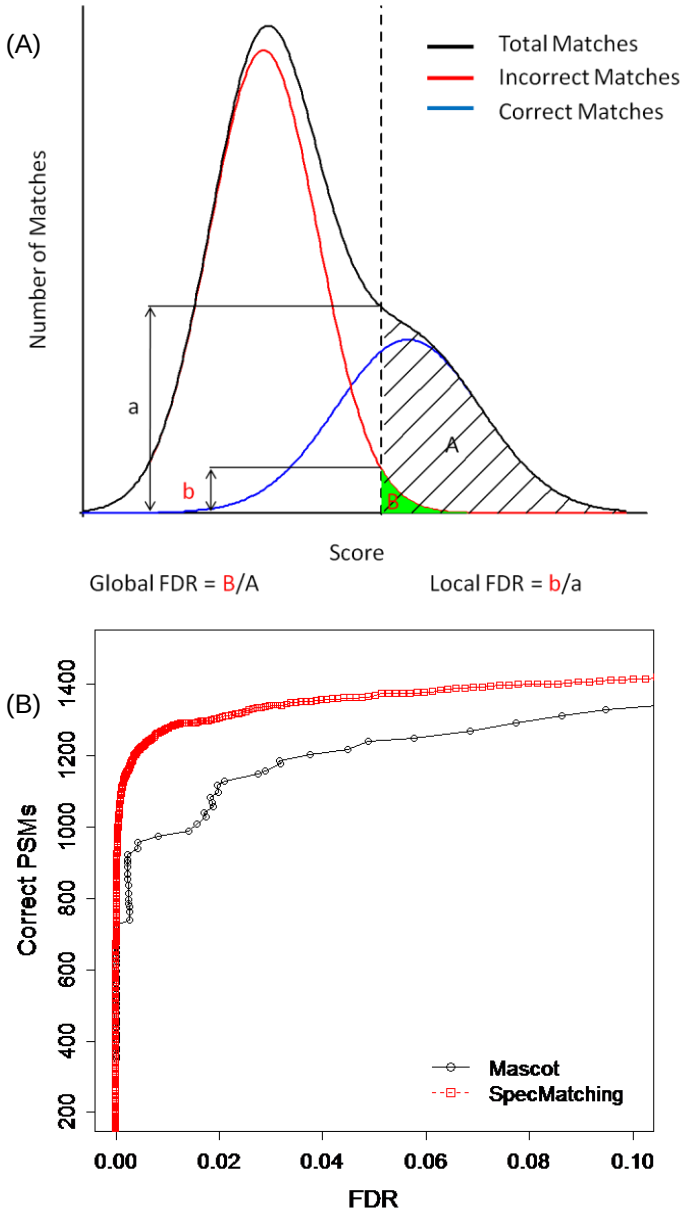

Figure 3.8 (A) Determination of the global and local false discovery rates. (B) Receiver-operating characteristics curves (ROC curves) of the search results obtained by Mascot search and SpecMatching spectral search.

The peptide matches that can only be identified by either search engines were then isolated and examined manually for their sequence assignments. Among the 97 matches only found in the Mascot result, 66 of them are due to the fact that their peptide sequences were not included in the spectral library. Again, by expanding the spectral library in the future, these matches should be found in the SpecMatching result. Twenty-eight of the 97 matches were low score incidences (lower than the Mascot cut-off score of an average of 30) in the SpecMatching result, while three of them were high-score second hits (higher than the score cut-off) in the SpecMatching result. After manually checking the spectra of the three incidences, it was suspected that SpecMatching failed to differentiate the correct from the incorrect PSMs, resulting in false negatives.

In the 270 matches found only in the SpecMatching result, 246 of them were identified by Mascot but rejected by identity threshold filter, while 24 of them could not be identified by Mascot. However, further analysis shows that 22 out of the 24 identifications had good retention time correlation with the peptide sequences stored in the spectral library; the average retention time difference of less than 1.5 min in a 2-h run was within the experimental error. This suggests that most of these matches were correct.

The above results indicate that SpecMatching is superior to Mascot searching in terms of both sensitivity and specificity. The major advantage mainly comes from the fact that the intensity pattern of the fragment ions in a spectrum is properly used in SpecMatching. This finding is consistent with the notion from a study comparing SpectraST with SEQUEST that spectral searching algorithm

worked better than sequence searching algorithm.[8] In our case, the correctness of the peptide spectra in the spectral library has been experimentally validated using the metabolic isotope labeling method.

3.4 Conclusions

A strategy has been developed to provide experimental evidence to validate the peptide match results generated from sequence-database searches of MS/MS spectra in shotgun proteome analysis. It is based on the use of metabolic isotope labeling to produce unlabeled and ^{15}N-labeled proteome samples from which tryptic peptides were produced for 2D-LC QTOF MS/MS analysis. The QTOF instrument offers relatively higher mass resolving power and mass measurement accuracy for MS/MS than other tandem MS with similar speed of spectral acquisition. The MS/MS spectra of the unlabeled peptides and their labeled counterparts can be overlaid and their fragmentation patterns and mass shifts due to nitrogen number differences can be readily compared to validate the spectrum-to-sequence matches. For spectral validation, two cut-off filters were developed. One was based on the number of common fragment ions found in the overlaid spectra; a minimum of 5 common ions were found to be needed to judge the fragmentation pattern matches. The second filter was based on the similarity of the fragmentation patterns of the unlabeled and labeled peptide pairs. A similarity score was calculated by using the fragment ion intensity dot-product,

and the cut-off score was found to be 0.96 out of 1.00, with 1.00 to be a perfect score.

Using *E. coli* K12 proteome analysis as an example, it has been shown that this strategy can be used to construct a more reliable MS/MS spectral library. By analyzing the whole cell lysate digests, a total of 257,907 and 245,156 spectra were acquired from the unlabeled and ^{15}N-labeled samples, respectively, using 2D-LC MS/MS. From the analysis of these spectra, an experimentally validated MS/MS spectral library of tryptic peptides was constructed. It consists of 9,302 unique spectra (unique sequence and charge state) from 7,763 unique peptide sequences. Finally, a spectral searching algorithm called SpecMatching was developed to utilize this spectral library. In analyzing a different digest of an *E. coli* extract using both Mascot and SpecMatching, it was shown that SpecMatching provided better sensitivity and specificity.

We envisage the use of this strategy to construct the MS/MS spectral library of various organisms for proteome analysis with improved sensitivity and specificity. To increase the number of validated spectra, it is necessary to develop an optimized peptide precursor ion inclusion strategy to generate more common spectra of the unlabeled and labeled counterparts. This strategy is demonstrated with details in Chapter 4. To generate a comprehensive MS/MS spectral library of a model organism, such as *E. coli*, more detailed proteome analysis, such as the use of cellular fractionation (e.g., membrane-bounded vs. plasma) and protein separation (e.g., based on molecular weights), followed by 2D-LC MS/MS, will

be needed. There are currently plans to construct a website containing spectral libraries of model organisms and a spectral search tool, including algorithms to address the spectral transferability issue related to the use of different MS/MS platforms, for shotgun proteome analysis.

3.5 Literature Cited

(1) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.

(3) Craig, R.; Cortens, J. P.; Beavis, R. C. *J. Proteome Res.* **2004**, *3*, 1234-1242.

(4) Nesvizhskii, A. I. *J. Proteomics* **2010**, *73*, 2092-2123.

(5) Yates, J. R., III; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557-3565.

(6) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. *J. Proteome Res.* **2006**, *5*, 1843-1849.

(7) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. *Anal. Chem.* **2006**, *78*, 5678-5684.

(8) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *Proteomics* **2007**, *7*, 655-667.

(9) Hummel, J.; Niemann, M.; Wienkoop, S.; Schulze, W.; Steinhauser, D.; Selbig, J.; Walther, D.; Weckwerth, W. *BMC Bioinformatics* **2007**, *8*, 8.

(10) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. *J. Proteome Res.* **2008**, *7*, 113-122.

(11) Falth, M.; Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Andren, P. E.; Zubarev, R. A. *J. Proteome Res.* **2007**, *6*, 4063-4067.

(12) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. *Nat. Methods* **2008**, *5*, 873-875.

(13) Ahrne, E.; Masselot, A.; Binz, P. A.; Muller, M.; Lisacek, F. *Proteomics* **2009**, *9*, 1731-1736.

(14) Zhang, X.; Li, Y. Z.; Shao, W. G.; Lam, H. *Proteomics* **2011**, *11*, 1075-1085.

(15) Lam, H.; Aebersold, R. In *Proteome Bioinformatics*; Hubbard, S. J., Jones, A. R., Eds.; Humana Press Inc, 2010, pp 95-103.

(16) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(17) Nelson, C. J.; Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *Proteomics* **2007**, *7*, 1279-1292.

(18) Cantin, G. T.; Venable, J. D.; Cociorva, D.; Yates, J. R., III *J. Proteome Res.* **2006**, *5*, 127-134.

(19) Dong, M.-Q.; Venable, J. D.; Au, N.; Xu, T.; Park, S. K.; Cociorva, D.; Johnson, J. R.; Dillin, A.; Yates, J. R., III *Science* **2007**, *317*, 660-663.

(20) McClatchy, D. B.; Dong, M.-Q.; Wu, C. C.; Venable, J. D.; Yates, J. R., III *J. Proteome Res.* **2007**, *6*, 2005-2010.

(21) Zhong, H.; Marcus, S. L.; Li, L. *J. Proteome Res.* **2004**, *3*, 1155-1163.

(22) Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* **2009**, *81*, 1049-1060.

(23) Liu, J.; Bell, A. W.; Bergeron, J. J. M.; Yanofsky, C. M.; Carrillo, B.; Beaudrie, C. E. H.; Kearney, R. E. *Proteome Sci.* **2007**, *5*, 12.

(24) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(25) Nelson, C. J.; Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *Proteomics* **2007**, *7*, 1279-1292.

(26) Snijders, A. P. L.; de Vos, M. G. J.; Wright, P. C. *J. Proteome Res.* **2005**, *4*, 578-585.

(27) Wang, N.; Li, L. *Anal. Chem.* **2008**, *80*, 4696-4710.

(28) Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. *Proteomics* **2005**, *5*, 3475-3490.

Chapter 4

Experimental Evaluation of Statistical Tools for Peptide and Protein Identification Using ^{18}O-labeling and Inclusion Strategy*

4.1 Introduction

In order to elucidate peptide sequences from mass spectrometry data, several strategies[1, 2] have emerged and matured. Amongst them, the most widely-used approach, sequence-database searching, involves comparing experimental MS/MS spectra with theoretical peptide fragmentation patterns derived from protein sequences in a proteome database and reporting the best peptide-spectrum matches (PSMs). Based on this concept, several sophisticated database search engines, such as Mascot[3], SEQUEST[4] and X!Tandem[5], have been developed. By evaluating the resultant matches using statistical tools, either individually or globally, a final list of peptide sequences can be generated according to a defined confidence level[6]. Even though these search engines share the same concept of calculating scores for PSMs and subsequently inferring protein identifications, they distinguish themselves in their individual ways of scoring and interpretation.

*A version of this chapter has been prepared for submission as Xu, M, Li, L., Experimental Evaluation of Statistical Tools for Peptide and Protein Identification Using ^{18}O-labeling and Inclusion Strategy.

In Mascot (Matrix Science, London, UK), a probability-based Mascot identity threshold for each individual PSM is reported. A Mascot ion score above the threshold is generally considered to be a significant peptide assignment. In the definition of Mascot identity threshold, $-10 \times \log_{10}(p/n)$, p is the defined error rate and n represents the number of candidate peptides (search space). For example, if there are 500 or 5000 candidate peptides and the probability of 0.05 is chosen, the Mascot identity threshold would be 40 or 50, respectively. Clearly, an increase in the number of candidate peptides, which is usually affected by searching in a more comprehensive database, enlarging the precursor tolerance windows, decreasing enzyme specificity, increasing the number of missed cleavages and variable modifications, would ultimately increase the Mascot identity threshold value. It is a well-known issue[7, 8] of Mascot scoring scheme, that Mascot identity threshold becomes very conservative (poor sensitivity) when dealing with a large search space.

Instead of reporting a probability and identity threshold for each PSM, X!Tandem[9] adopted the concept of reporting expectation values (E-values) of PSMs. Unlike Mascot, X!Tandem is an open source software. In its algorithm, X!Tandem first measures the spectral similarity between the experimental spectrum and several candidate theoretical peptide fragmentation patterns, generates hyperscores (the sum of matched fragment ion intensities multiplied by the N factorial for the number of matched ions), plots a distribution of hyperscores for the spectral search and extrapolates an E-value to provide a statistical evaluation for each identification. E-value is defined as the number of

random matches that would be expected to have the same or better scores. It is believed that this empirical scoring scheme is more sensitive when searching a large space. Mascot also implemented this idea in its software package, called Mascot homology threshold. However, due to their different ways of performing statistical analysis, E-values from X!Tandem and probabilities from Mascot are not directly comparable.

In a recent development, researchers have focused more on the statistics used to evaluate the resulting PSMs, instead of sequence-database algorithms themselves. The target-decoy strategy was developed to estimate the global false-discovery rate (FDR) of MS/MS search results. In the target-decoy approach, the target database represents the normal proteome database, and the decoy database is a reversed or randomized version of the target database. Any PSM from the decoy search that passes the score threshold is deemed a false positive. Based on the number of estimated false positives, the global FDR can be readily estimated. As the most commonly used approach, some researchers prefer searching the target and decoy databases separately, while others prefer constructing a concatenated database by combining the target and decoy databases and searching this compound target/decoy database. In 2007, Elias and Gygi[10] published a detailed study of various target-decoy strategies.

Apart from a simple target-decoy strategy, more sophisticated algorithms were developed to re-evaluate match scores and assign probabilities to each PSM by examining the properties of correct and incorrect PSMs. For example,

PeptideProphet[11] takes advantage of the bimodal distribution that discriminant scores of correct and incorrect PSMs in the histogram and uses an expectation-maximization algorithm to fit the distribution and thus calculate probabilities of each PSM. On the other hand, Percolator[12] adopts a different machine learning approach. After extracting a vector of features that are related to the quality of the match (e.g., mass error and PSM score) from both target and decoy PSMs, an iterative classification process is applied to discriminate the target and decoy PSMs by using those features. After several iterations, the system converges and generates a robust classifier that can be used to calculate the probability of each PSM being a random match. Both methods have shown to be able to achieve high sensitivity without sacrificing too much specificity.[11,12] While Percolator only applies to SEQUEST and Mascot search results[13], PeptideProphet has been adapted for X!Tandem search results as well.

Despite these significant advances in statistical evaluation of PSMs, statistical tools do not provide experimental validation of the spectrum-to-sequence assignments. As shown in Chapter 3, by using differential isotope labeling, experimental evidence can be provided to validate protein identification results generated by sequence-database search method. In this Chapter, an ^{18}O-labeling approach was applied to complex peptide mixtures after trypsin digestion of protein mixtures. Traditionally, ^{18}O-labeling methods have been applied to detect differentially expressed proteins in various proteomic systems[14, 15]. In the case of trypsin-catalyzed ^{18}O-labeling, two carboxyl-terminal ^{16}O atoms are substituted with two ^{18}O atoms in ^{18}O-enriched H_2O medium. The resultant

labeled peptides would have a higher molecular mass (4.0085 Da higher) than their unlabeled counterparts, but still behave the same way during separation in reversed-phase chromatography, as well as during the ionization and fragmentation processes in a mass spectrometer. Therefore, it is easy to detect and isolate the unlabeled and labeled pair in RP LC-MS/MS runs simply based on their retention time and precursor mass difference. Previous studies have demonstrated that, by examining the unlabeled and labeled MS/MS spectra of the same peptide sequence, evidence can be provided to validate search results generated by *de novo* sequencing[16-18], peptide mass fingerprinting[19], as well as the sequence-database strategy[20, 21].

Herein an approach is developed involving the use of ^{18}O-labeling for validating the PSMs generated from sequence-database searches using both Mascot and X!Tandem. Advanced statistical tools, including PeptideProphet and Percolator, were also assessed. In the experimental workflow, an inclusion strategy enabled targeted analysis for ^{18}O-labeled PSMs in the LC-MS/MS runs. With multiple inclusion runs, almost all pre-identified unlabeled PSMs had matching ^{18}O-labeled counterparts found. A retention time-based data filtering strategy enabled the isolation of true and false identifications for all the unlabeled PSMs. Based on these experimentally validated PSMs, the performance of all the statistical tools was carefully inspected.

4.2 Experimental Section

4.2.1 Chemicals and Reagents

Dithiothreitol (DTT), iodoacetamide (IAA), trifluoroacetic acid (TFA), guanidinium hydrochloride and ammonium bicarbonate were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). HPLC grade formic acid, LC-MS grade water, acetone, and acetonitrile (ACN) were from Fisher Scientific Canada (Edmonton, Canada). The BCA assay kit and immobilized trypsin were purchased from Pierce (Rockford, IL).

4.2.2 Sample Preparation

SU-DH-L1 cells[22] (A human lymphoma cell line) were cultured in Dulbecco's modified Eagle's medium (Sigma) supplemented with 10% heat-inactivated fetal bovine serum and 1% penicillin under an atmosphere of 95% O_2 and 5% CO_2 in 98% humidity at 37 ℃. Cells were collected and lysed in CelLytic^{TM} M buffer (Sigma), 1mM phenylmethylsulfonyl fluoride (Sigma) and protease inhibitor mixture (Sigma). The lysates were precleared at 20,000 *g* for 15 min. BCA assay on the cell lysate solution was performed to determine the protein concentration. Proteins in the cell lysates were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide (IAA). Acetone pre-cooled to -80 ˚C was added gradually to the cell lysates to a final concentration of 80% (v/v). The solution was then incubated at -20 ˚C overnight and centrifuged at 20,000 *g* for 15 min. The supernatant was decanted and the pellet was carefully washed once

using cold acetone. The pellet was re-solubilized in 6 M guanidinium hydrochloride. The protein mixtures were then fractionated by reversed-phase liquid chromatography on an Agilent 1100 HPLC system (Palo Alto, CA) using a 4.6 X 50 mm mRP-C18 High-Recovery Protein Column (Agilent, CA) at 75 ˚C. In total, 40 fractions were collected. Next, each protein fraction was then dried down using a SpeedVac concentrator system and reconstituted using 50 mM ammonium bicarbonate. Pre-treated immobilized trypsin (50 µL) was added to each fraction, followed by 24 hours of incubation at 37 ˚C with rapid shaking. The immobilized trypsin gel in each fraction was then separated from the digestion mixture using a resin separator (Pierce). Desalting of each peptide fraction was performed on a 4.6 mm × 50 mm Polaris C18 A column with 3 µm particles and 300 Å pore size (Varian, CA). The eluted peptides were monitored and quantified using a UV detector operated at 214 nm. Each fraction then was divided into two equal portions, one of which underwent the ^{18}O-labeling process. In the labeling process, both portions of each fraction were dried down. Then one portion was reconstituted in 50 µL of ^{18}O-enriched H_2O (Cambridge Isotope Laboratories, Andover, MA), followed by the addition of 50 µL of immobilized trypsin to accelerate the carboxyl-terminal oxygen exchange. The other portion underwent the same process except normal H_2O was used instead of ^{18}O-enriched H_2O. After 24 hours of incubation at 37 ˚C, immobilized trypsin was removed from the peptide mixture using a resin separator. Unlabeled and ^{18}O-labeled samples were then kept separately at -80 ˚C until further analysis.

4.2.3 RPLC MS/MS

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, 1 μg of the digest was injected onto a 75 μm × 100 mm Atlantis dC18 column with 3 μm particles and 100 Å pore size (Waters, Milford, MA) via a Symmetry C18 trap column (180 μm × 20 mm). For the chromatographic separation, solvent A consisted of 0.1% formic acid in water and solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using a 120-min gradient and introduced by electrospray into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants (i.e., lock-masses), was infused at a flow rate of 300 nL/min, and a 1-s MS scan was acquired every 1 min throughout the run.

4.2.4 Inclusion Strategy

As shown in Figure 4.1, the unlabeled and ^{18}O-labeled aliquot of the same fraction were first analyzed in ordinary RPLC-MS/MS experiments to identify the unlabeled and ^{18}O-labeled peptides, respectively, as described in the RPLC MS/MS section. Then the peptide identification lists from both aliquots were compared to isolate the common peptide sequences and the unique peptide sequences from the unlabeled aliquot. Since ^{18}O-labeling approach increases the molecular masses of all the tryptic peptides with C-terminal lysine or arginine by 4.0085 Da, it is possible to calculate the theoretical *m/z* values of those peptides when they are ^{18}O-labeled. In principle, due to their identical chromatographic

and fragmentation behavior, in an ^{18}O-labeled run of the same fraction the ^{18}O-labeled peptides would have exactly the same retention times as their unlabeled counterparts in their unlabeled runs. Therefore, based on the calculated *m/z* and its corresponding retention time, an inclusion list can be generated for the following ^{18}O-labeled run. In the following ^{18}O-labeled run, inclusion settings are switched on to make sure that the mass spectrometer only analyzes the ions with the *m/z* values and retention time specified on the inclusion list. Then all the ^{18}O-labeled peptides were compared with the unlabeled peptide identifications to isolate common peptide sequences and the unique peptides from the unlabeled aliquot. Ideally, after one inclusion run, all the true positives from the unlabeled aliquot will find their ^{18}O-labeled counterparts and the ones that do not have ^{18}O-labeled counterparts identified are highly likely to be false identifications. However, due to imperfect reproducibility of the liquid chromatography and ionization process, one inclusion run is insufficient to find all the ^{18}O-labeled counterparts for all the unlabeled peptides. Therefore, multiple inclusion runs are needed to resolve this issue.

In order to estimate how many inclusion runs are needed, a replicate study was carried out. In this study, an unfractionated complex human cell lysate sample was prepared. In a similar manner to the inclusion strategy shown in Figure 4.1, a technical replicate of the unlabeled sample was used to replace the ^{18}O-labeled fraction. Instead of calculating ^{18}O-labeled precursor *m/z* values, the unlabeled precursor *m/z* values were used in the inclusion runs. Therefore, the minimal number of inclusion runs that are needed for ^{18}O-labeled inclusion can be

readily estimated by how many technical replicates are needed to re-analyze all the unlabeled peptide identifications by using the same strategy. It was found that 5 LC-MS/MS runs (1 normal LC-MS/MS run and 4 inclusion LC-MS/MS runs) were the minimal requirement for an ^{18}O-labeled inclusion study (see results and discussion section 4.3.1). In total, there were 40 unlabeled fractions from protein RPLC fractionation and 253 ^{18}O-labeled runs were analyzed. The average number of ^{18}O-labeled runs per unlabeled fraction was 6.325.

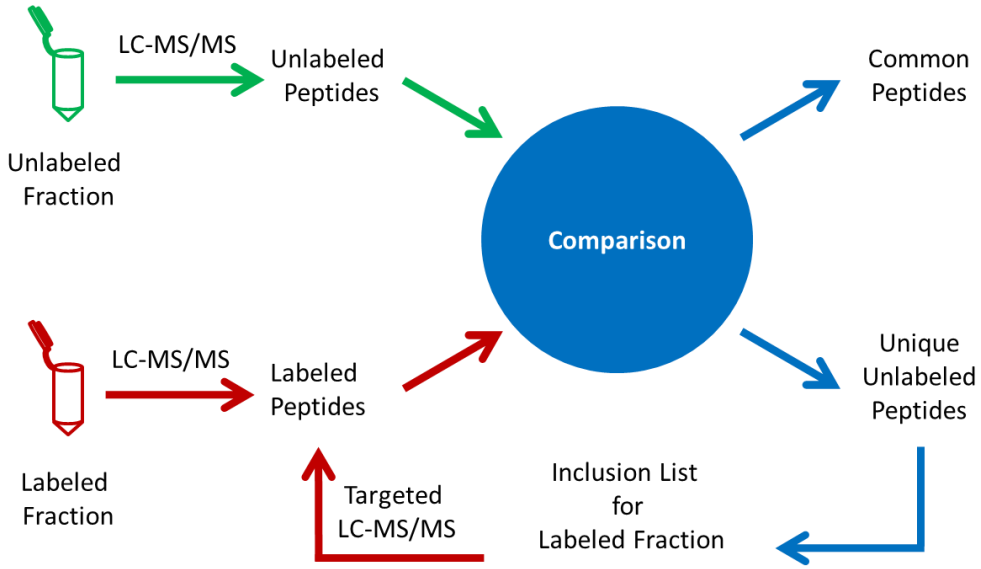

Figure 4.1 Schematic of inclusion strategy.

4.2.5 Mascot Search

Using Proteinlynx Global Server 2.3.0 (Waters) all raw LC-ESI data were lock-mass corrected, de-isotoped, and converted to peak list files with retention time information. All the peak list files were then submitted to the Mascot search program (version 2.2.1). The search parameters for unlabeled samples were

selected as follows: enzyme, trypsin; missed cleavages, 1; peptide tolerance, 30 ppm; MS/MS tolerance, 0.2 Da; fixed modification, carbamidomethyl (C); variable modifications, ammonia-loss (N-term C), N-Acetyl (protein), oxidation (M), pyro-Glu (N-term Q), and pyro-Glu (N-term E). The search parameters for the ^{18}O-labeled samples were the same as the unlabeled samples with one additional variable modification, double ^{18}O-modification of carboxyl-terminal lysine or arginine. A concatenated database was constructed by combining the IPI Human database (version 3.68) and its reversed proteome sequences. For Mascot Percolator results, database searching was restricted to the original IPI Human database (version 3.68) with decoy function switched on in the Mascot program. For Mascot PeptideProphet results, database searching was restricted to the concatenated database.

4.2.6 X!Tandem Search

All the peak list files were also submitted to X!Tandem search engine (The Global Proteome Machine Organization, 2007.07.01). For the purpose of fair comparison, in all the X!Tandem searches the search parameters, including the variable modification settings in the refinement function, were kept the same as Mascot searches. The search was restricted to the aforementioned concatenated database.

4.2.7 Statistical Analysis

In the original Mascot search results, a significance threshold of 0.05 was applied to identify peptide sequences from each LC-MS/MS run. If a peptide-

spectrum match (PSM) has a Mascot ion score no less than its Mascot identity threshold, it was deemed as identified. However, X!Tandem implements a different scoring scheme than Mascot. In X!Tandem, an E-value (Expectation value) is calculated for each PSM to provide statistical evaluation. In its definition[9], E-value is the number of random matches that would be expected to have the same or better scores. Therefore, the higher the E-value, the less likely a PSM is deemed to be a valid identification. It was observed that by applying a maximum E-value of 0.05 to X!Tandem results, the estimated global false-discovery rates of Mascot and X!Tandem results for the same LC-MS/MS run were very similar. Therefore, in the original X!Tandem search result, an arbitrary maximum E-value for PSM was set to 0.05.

The target-decoy search strategy proposed by Elias and Gygi in 2007[10, 23] was applied by searching the MS/MS spectra against the concatenated database to calculate the global false discovery rate (FDR). The number of false positives can be estimated by counting the number of decoy sequences above the score threshold. The estimated global FDRs were calculated by the number of false positives divided by the total number of identifications (false positives + true positives).

Mascot Percolator[13] was also used to statistically evaluate the Mascot result in order to improve the number of identifications. In this study, the default setting was used, including all the features. After being processed by Percolator, each PSM was assigned two statistical values, posterior error probability (PEP)

and q-value. The q-value of each PSM can be understood as the minimal global FDR that is required to include that PSM in the search result[12, 24]. PEP can be deemed as the local FDR of a PSM, which indicates the probability of such match being random. In Mascot Percolator processed results, the original Mascot p-values are replaced with PEP value. Therefore, based on the Mascot scoring equation, Mascot Score = $-10 \times \log_{10}(p)$, the new Mascot score of 13 represents the local FDR of such PSM is 0.05. By applying the Mascot score threshold of 13 to the Mascot Percolator result, it is guaranteed that the maximum local FDR of the search result is 5%. In this study, the minimum new Mascot score of 13 was chosen.

PeptideProphet[11, 25] was the other machine learning algorithm used to statistically evaluate the results both from Mascot and X!Tandem. As a part of statistical software called Trans Proteomic Pipeline (version 4.3, JETSTREAM REV 1, Build 200909091257 MinGW), it assigns a probability to each PSM. Based on the definition of PeptideProphet probability, it is the probability of a PSM being correct. Therefore, 1 – probability can be understood as the local FDR of a PSM. In this study, a concatenated search and non-parametric modeling[26-28] were used. Using this strategy the negative distribution can be readily pinned down by decoy hits when fitting the bimodal distribution in PeptideProphet to maximize peptide identification while maintaining a low error rate. In PeptideProphet-processed results, the minimum probability of 0.95 (local FDR of 0.05) was chosen.

4.2.8 ^{18}O-labeling Validation

A straightforward workflow was devised to experimentally validate unlabeled PSMs. First of all, all the unlabeled PSMs had to be confidently identified by at least one of the statistical tools: Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem or X!Tandem PeptideProphet. For example, in Mascot an unlabeled PSM had to have an ion score higher than its identity threshold when a significance threshold of 0.05 was applied. PSMs of lower confidence were not included in the downstream analysis.

Secondly, all the ^{18}O-labeled counterparts of the unlabeled PSMs had to be confidently identified by at least one of the aforementioned statistical tools as well. It ensures the reliability of the labeled PSMs used to validate the unlabeled PSMs. For example, if both the unlabeled and labeled identification of the same peptide sequence were identified by Mascot PeptideProphet with a probability of 0.95, the chance of neither PSM being incorrect is extremely low (0.25%). Therefore it is of great importance to obtain good quality ^{18}O-labeled PSMs.

Lastly, both unlabeled and labeled identifications of the same peptide sequence had to elute at a similar organic solvent composition (%B = initial %B + (retention time – dead time) × gradient slope) during a RPLC separation. Theoretically, if both identifications were correct, they would have the same chromatographic behavior. Thus ideally, they would have identical %B when the same LC gradient profile is used. In reality, however, each peptide elutes from a chromatographic column as a Gaussian shape peak. Consequently, the same

peptide sequence might be associated with slightly different %B each time. So in this study, the last validation filter was the %B difference between the unlabeled and labeled peptides of the same sequence. If they differed too much, the unlabeled PSM was not considered to be validated by that labeled PSM, even if they appeared to share the same sequence.

Only when it fulfilled all three validation requirements, was an unlabeled PSM deemed validated.

4.2.9 Data Processing

All in-house programs were written in Perl 5.12 (http://www.perl.org). Charts and graphs were generated using R's plotting packages (http://www.r-project.org/) and Microsoft Excel 2007. Software was run on standard desktop and laptop computers running Windows 7 (Home Edition).

4.3 Results and Discussion

4.3.1 Inclusion Strategy

First of all, it is necessary to demonstrate that the proposed inclusion strategy could effectively re-identify PSMs found in the original LC-MS/MS run. As described in the experimental session, an unfractionated complex human cell lysate sample was prepared and analyzed by RPLC-MS/MS. After extracting all the MS/MS spectra, searching them using different search engines (e.g., Mascot and X!Tandem) and analyzing the results with various statistical tools (e.g.,

PeptideProphet and Percolator), an original identification result, listed in Table 4.1, was obtained. As shown in Table 4.1, 1601 PSMs were found in the Mascot result, in which only the original Mascot significance threshold of 0.05 was applied. The estimated global FDR of the Mascot result is 0.3%. Furthermore, PeptideProphet was applied to this Mascot search result to provide an advanced statistical analysis of those peptide identifications. It was found that by applying a minimal PeptideProphet probability of 0.95, which can be understood as the maximal local FDR of 0.05, 2587 PSMs were found with an estimated global FDR of 0.6%. In the meanwhile, Mascot Percolator was applied to the Mascot search result to provide a different statistical analysis. With a maximum posterior error probability (PEP) of 0.05, 2605 PSMs were found with an equivalent q-value cut-off of 0.4%. Similarly, X!Tandem was also used to search all the MS/MS spectra in the same human proteome database. After applying an arbitrary maximum E-value of 0.05 to the search result, 2070 PSMs were identified with a global FDR of 0.3%. Then PeptideProphet was also applied to X!Tandem results to provide a different statistical analysis. By applying a maximal local FDR of 0.05, 2402 PSMs were found with a global FDR of 0.4%. Then all the identified PSMs from all the aforementioned statistical tools were combined to construct a target list in order to apply inclusion strategy. After 5 inclusion runs, almost all the PSMs from the original run were able to be re-identified. As shown in Table 4.1, only 5 PSMs out of 1601 were not re-identified in the Mascot result. The inclusion percentage was as high as 99.7%. It was found that more inclusion runs did not further improve the inclusion percentage. Considering the fact that the

Mascot result of the original run was analyzed with a significance threshold of 0.05 and the estimated global FDR was 0.3%, it is reasonable to state that all the correct PSMs were effectively re-identified using the inclusion strategy. Similarly high inclusion percentages were also observed in the other search results: Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet (see Table 4.1).

Table 4.1 Inclusion Strategy Results.

	Mascot	Mascot PeptideProphet	Mascot Percolator	X!Tandem	X!Tandem PeptideProphet
Total PSMs	1601	2587	2605	2070	2402
Leftover PSMs[a]	5	57	55	36	65
Inclusion Rate	99.7%	97.8%	97.9%	98.3%	97.3%

a: Leftover PSMs are those PSMs found in the original run but cannot be re-identified in the inclusion runs.

4.3.2 Identification Result Summary

Table 4.2 summarizes the results obtained from the RPLC-QTOF MS/MS analysis of the human anaplastic large cell lymphoma cell line (SU-DH-L1). In total, 401,762 and 571,675 spectra were collected from the unlabeled and ^{18}O-labeled samples, respectively. All the spectra were searched by both Mascot and X!Tandem search engines and all the potential PSMs were given scores or probabilities after being statistically evaluated by Mascot, PeptideProphet, Mascot Percolator and X!Tandem. Arbitrary score thresholds were set up for all the statistical tools. In the original Mascot, a significance threshold of 0.05 was used. For PeptideProphet results, a minimal probability of 0.95 was applied. For Mascot

Percolator results, the new Mascot Identity threshold of 13, which is equivalent to a maximum posterior error probability (PEP) of 0.05, was applied. In the original X!Tandem results, the maximum expect value of 0.05 was adopted to maintain a similar global FDR as the original Mascot result where the significance threshold of 0.05 was applied.

As shown in Table 4.2, 94,121 unlabeled PSMs, corresponding to 3,263 proteins, were identified by the original Mascot search engine with an estimated global FDR of 0.4%. Among those PSMs, 93,381 are matches with C-terminal lysine or arginine. A new term, KR PSM, was coined to represent those PSMs. Theoretically only KR PSMs can be labeled by ^{18}O-labeling process. After removing redundant PSMs, 19,123 unique PSMs were obtained. Instead of only processing the Mascot result from the unlabeled data set by Mascot, it was found that after evaluating Mascot results by PeptideProphet or Mascot Percolator, the total numbers of PSMs was evidently higher at 120,105 and 130,774, increasing the original Mascot result by 27.6% and 38.9%, respectively. Consequently, the numbers of proteins identified by Mascot PeptideProphet and Mascot Percolator were also higher: 4,130 and 3,946, respectively. The Mascot PeptideProphet result still maintained a relatively low global FDR of 0.5% and the Mascot Percolator result had a low q-value score of 0.4%, indicating an error rate similar to that of the original Mascot result. When processing the unlabeled spectra by the X!Tandem search engine, instead of Mascot, 99,186 unlabeled PSMs were obtained. Among them, 98,175 could be categorized as KR PSMs. After removing redundant PSMs, 19,671 unique PSMs were found. After processing the

X!Tandem results with PeptideProphet, a 10.1% improvement in PSMs was observed. Global FDR estimation was also performed to ensure a good quality of data. It was found that the estimated global FDR was 0.4% in both X!Tandem and X!Tandem PeptideProphet results. When all the results from different statistical tools were combined, a total of 141,389 unlabeled PSMs, including 140,079 were KR PSMs, were found.

Table 4.2 Identification Result Summary.

	Unlabeled Data Set				^{18}O-labeled Data Set			
	All the PSMs	KR PSMs[a]	Unique PSMs[b]	Proteins[c]	All the PSMs	KR PSMs[a]	Unique PSMs[b]	Proteins[c]
Mascot	94,121	93,381	19,123	3,263	125,540	124,308	21,682	3,590
Mascot PeptideProphet	120,105	119,105	23,941	4,130	178,642	176,854	28,854	4,766
Mascot Percolator	130,774	129,634	25,155	3,946	192,932	191,024	30,020	4,481
X!Tandem	99,186	98,175	19,671	3,498	149,571	147,852	23,861	4,115
X!Tandem PeptideProphet	109,229	108,150	21,357	3,623	166,493	164,616	25,786	4,265
Combined Result	141,389	140,079	27,740	4,982	211,584	209,298	33,154	5,661

a. KR PSMs are the PSMs that contain C-terminal lysine or arginine. Only KR PSMs can be labeled by ^{18}O-H_2O in this experiment.

b. Unique PSMs are calculated by removing the redundant PSMs from all the PSMs;

c. Protein identifications are derived from all the PSMs, not just K, R PSMs.

Detailed analysis showed that 55.6% of all the PSMs were identified by all five statistical tools (see Figure 4.2), whilst each statistical tool has its own strength in matching peptide sequences to MS/MS spectra. Of them, the original Mascot results had the lowest number of identifications, due to the fact that Mascot rejects many PSMs because of the high identity thresholds with Mascot (an average identity threshold of 33). It is a well-known issue of the original Mascot scoring scheme that the Mascot identity threshold is unreasonably high for relaxed parameter settings or when searching very large databases[8]. Meanwhile, Mascot Percolator had the highest number of PSMs, covering 97.7% of the original Mascot result and 94.3% of the Mascot PeptideProphet result. It clearly showed the high sensitivity of Mascot Percolator in terms of identifying PSMs. It is believed that the main improvement of Mascot Percolator comes from its advanced statistical analysis, with which it can discern many false negative identifications with relatively low Mascot ion scores from random matches.

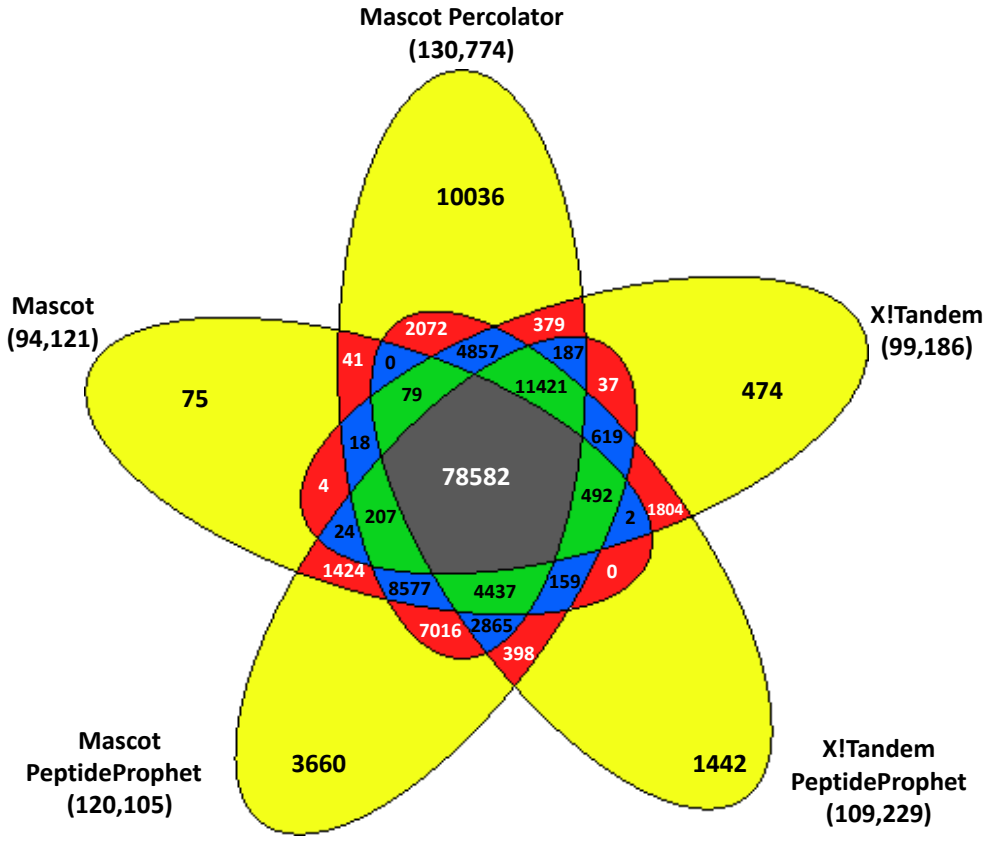

Figure 4.2 Venn diagram analysis of all the unlabeled PSMs from Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet.

Comparison of the original Mascot result with the original X!Tandem result shows that 79,408 PSMs could be identified by both search engines, constituting 84.4% of the original Mascot result and 80.1% of the X!Tandem result. The identification differences mainly originate in the different statistical analyses those two search engines provide. As shown in Figure 4.2, when PeptideProphet was applied to both search results to provide similar statistical analyses, the number of common PSMs increased by 19,565 (24.6%) to a total of 98,973. It indicated that both search engines often assign the same sequence to the same spectrum but different confidence levels to the same PSM. As a

result, some PSMs that are agreed upon by both search engines, especially the ones with confidence levels close to the cut-off value, might not pass the score thresholds of those two search engines at the same time to be considered as identified. While both search engines agree in most cases, X!Tandem can still identify PSMs that Mascot cannot (Figure 4.2), and vice versa. However, the validity of those PSMs, as well as all the common PSMs, remains to be examined.

After applying the inclusion strategy to the ^{18}O-labeled data set, 125,540, 178,642, 192,932, 149,571 and 166,493 labeled PSMs were identified with estimated global FDRs of 0.3%, 0.6%, 0.4% (q-value), 0.4% and 0.4%, from Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet, respectively. After combining all the PSMs, 211,584 PSMs were found. Upon removal of all the redundant PSMs, 33,154 unique PSMs remained. Because of the inclusion strategy applied in the experiment, a high overlap between the unlabeled and labeled data sets was to be expected. In fact, 96.5% of PSMs in the unlabeled data set had ^{18}O-labeled counterparts.

In summary, upon the protein level fractionation and RPLC-QTOF MS/MS analysis of the human cell lysates, a large collection of unlabeled and labeled PSMs were identified with relatively low estimated global FDRs.

Clearly, with the assistance of advanced statistical tools (PeptideProphet and Percolator), the number of identifications can be significantly improved. However, the validity of those PSMs remained questionable due to the lack of experimental corroboration. Therefore an ^{18}O-labeling validation method was designed to address this issue. In principle, all the correct unlabeled PSMs that had C-terminal lysine or arginine (KR PSMs) could be experimentally corroborated by their labeled counterparts, while it was unlikely for random matches to find a labeled counterpart and thus become validated.

4.3.3 ^{18}O-labeling Validation

As described in the experimental section, all the KR PSMs from the unlabeled data set had undergone ^{18}O-labeling validation. In brief, during the process of validation, an unlabeled PSM needed to pass three filters: 1. the unlabeled PSM needed to be confidently identified by at least one of the statistical tools; 2. one confidently identified ^{18}O-labeled counterpart had to be found; 3. both unlabeled and labeled PSMs of the same peptide sequence had to elute at similar %B.

In total, 3,914,793 comparisons were done, including comparisons between redundant peptide identifications. In the process of validation, the first two filters can be easily set up. As described in the experimental session, a significance threshold of 0.05 was used for the original Mascot results, while

a minimum probability of 0.95 was adopted for all the PeptideProphet results. At the same time a minimum new Mascot score of 13 (equivalent to maximum local FDR of 0.05) was chosen for Mascot Percolator results. Finally a maximum E-Value of 0.05 was selected for X!Tandem results. By implementing those two quality filters, it was ensured that most low quality PSMs were excluded from the validation process. Consequently, the chance of erroneously validating by low quality PSMs was greatly reduced.

Lastly, the third quality filter was set up and applied. If an unlabeled and a labeled PSM lead to the same peptide sequence identification, they constitute a spectral pair. Theoretically, in a spectral pair, if both PSMs were true identifications, they would elute at very similar %B from a chromatographic column. Therefore, it is reasonable to state that if in a spectral pair the %B at which the unlabeled and labeled PSMs elute differ too much, it is very likely that at least one of them does not contain the correct peptide identification information. In order to determine a reasonable %B difference cut-off between labeled and unlabeled PSMs in a spectral pair, a control experiment was carried out by analyzing the labeled sample in replicate and then examining the %B difference between replicate identifications. Figure 4.3A shows the distribution plot of the %B difference between replicate PSMs. It is very similar to Figure 4.3b, which is the distribution plot of the %B difference between labeled and unlabeled PSMs in

spectral pairs. As illustrated in both figures, most comparisons present a %B difference within 0.5%. The distribution shown in Figure 4.3B reveals that 95% of the cases have a %B difference of 0.5% or lower. Thus, a cut-off of 0.5% was adopted as the maximum %B difference to reject potentially false peptide matches during comparisons of unlabeled and labeled peptide spectra.

After applying those three quality filters, true positives can be isolated from all the identifications in each statistical result. Initial analysis showed that KR PSMs from the original Mascot result only constituted 66.7% of total KR PSMs, while the percentages for the Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet results were 85.0%, 92.2%, 70.1% and 77.2%, respectively (see Table 4.3). These results clearly showed that Mascot Percolator had the highest sensitivity in terms of identifying PSMs. After validation, a total of 136,027 unlabeled PSMs (3,954 proteins) could be validated by this strategy, which constituted 97.1% of the total unlabeled KR PSMs. As shown in Table 4.3, Mascot Percolator provided the highest number of KR PSMs (129,634), constituting 92.2% of the total KR PSMs identified, as well as the highest number of validated PSMs (127,405), constituting 93.7% of the total validated PSMs. At the same time, the validation rate of the Mascot Percolator result maintained a high value (98.3%), comparable to the validation rates from the other statistical tools, 99.6%, 98.5%, 99.2% and 99.0% for Mascot, Mascot PeptideProphet,

X!Tandem and X!Tandem PeptideProphet, respectively. Clearly, Mascot Percolator is not just the most sensitive tool to identify PSMs, but also provides fairly reliable results.

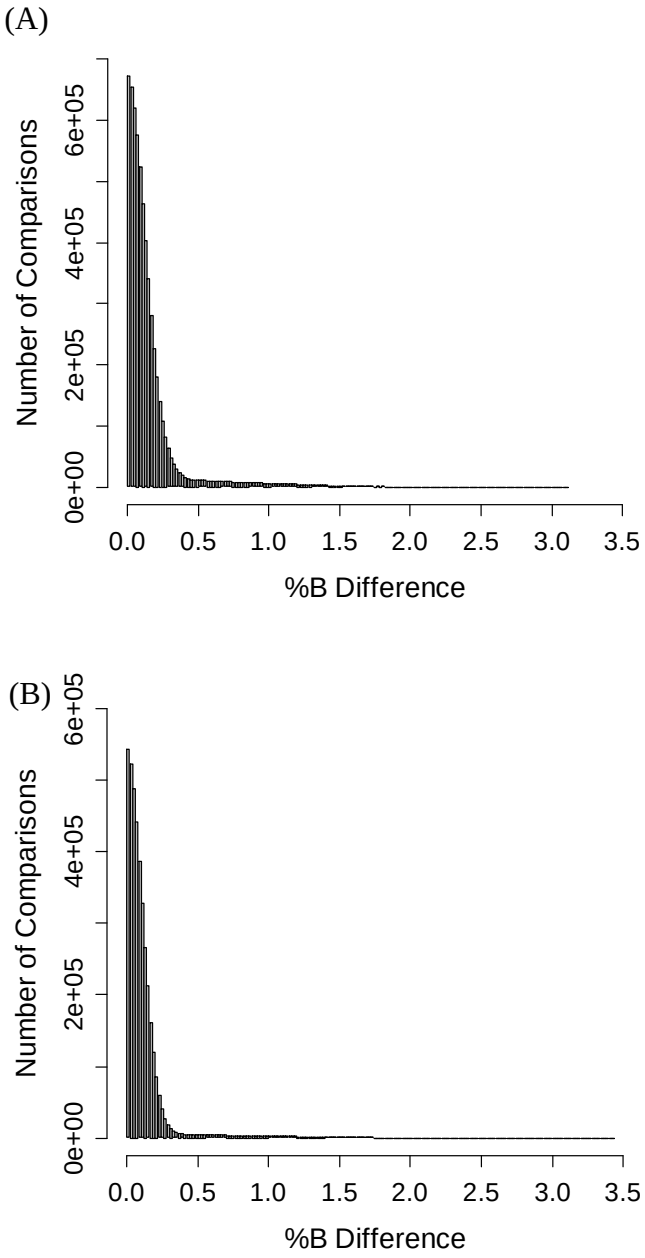

Figure 4.3 Number of comparisons as a function of eluting organic solvent composition (%B) difference from (A) the comparison of replicate identifications and (B) the comparison of unlabeled and labeled matches.

Table 4.3 Validation Summary.

Tools	KR PSMs[a]	Fractions of KR PSMs[b]	KR Proteins[c]	Validated PSMs	Validation Rate[d]	Fractions of Validated PSMs[e]	Validated Proteins
Mascot	93,381	66.7%	3,256	93,041	99.6%	68.4%	3,180
Mascot PeptideProphet	119,105	85.0%	4,114	117,363	98.5%	86.3%	3,612
Mascot Percolator	129,634	92.2%	3,933	127,405	98.3%	93.7%	3,618
X!Tandem	98,175	70.1%	3,476	97,364	99.2%	71.6%	3,285
X!Tandem PeptideProphet	108,150	77.2%	3,607	107,108	99.0%	78.7%	3,386
Combined Result	140,079	100%	4,944	136,027	97.1%	100%	3,954

a. KR PSMs are the PSMs that contain C-terminal lysine or arginine. Other tryptic PSMs are not included due to their inability of being labeled by the ^{18}O strategy.

b. Fractions of KR PSMs are calculated by dividing the number of KR PSMs from one statistical tool by the total number of KR PSMs from all tools.

c. KR protein identifications are derived from KR PSMs.

d. Validation Rate is calculated by dividing the number of validated KR PSMs from one statistical tool by the total number of KR PSMs from the same tool.

e. Fractions of KR PSMs are calculated by dividing the number of validated KR PSMs from one statistical tool by the total number of validated KR PSMs from all tools.

Next, PSMs that are identified only by one statistical tool were examined to assess their chance of being validated. It is commonly believed that if a PSM can be identified by multiple tools, it is considered more reliable than the one that can only be identified by one tool. As shown in the Venn diagram analysis (see Figure 4.2), in Mascot Percolator results, there are 10,036 PSMs that are only found by Mascot Percolator. The numbers for the other statistical tools used in this study are 75, 3,660, 474 and 1,442 for Mascot, Mascot PeptideProphet, X!Tandem and X!Tandem PeptideProphet, respectively. Among all those tools, Mascot Percolator has the highest number of tool-specific PSMs. However, the question of their reliability still remains unanswered. The validation rates of those tool-specific PSMs were calculated after the application of the ^{18}O-labeling validation method. Figure 4.4, a bar graph of these validation rates shows that Mascot-specific PSMs had the lowest validation rate. However, considering the low number of Mascot-specific PSMs, no convincing conclusion can be drawn. However, for all the other tools where more than 400 tool-specific PSMs were found, it is statistically meaningful to calculate validation rates. As shown in Figure 4.4, Mascot Percolator-specific PSMs have the highest validation rate (88.5%), compared with the validation rates of 76.0%, 56.2% and 82.2% in Mascot PeptideProphet-specific PSMs, X!Tandem-specific PSMs and X!Tandem PeptideProphet-specific PSMs, respectively. Compared with the overall PSMs

validation rate (97.1%), they are still noticeably lower, indicating that, in general, those tool-specific PSMs are not as reliable as the ones that can be identified by multiple tools. Further study showed that PSMs identified by more statistical tools are more likely to be validated, thus being true PSMs. In Figure 4.5, the PSMs' validation rate is plotted against the number of tools that they can be found in. Clearly, as the number of tools increases, the validation rate of PSMs, which indicates that the reliability of PSMs, increases as well. In summary, this study experimentally corroborated the common belief that if PSMs can be identified by multiple tools, they are more reliable than those that can only be identified by one tool.

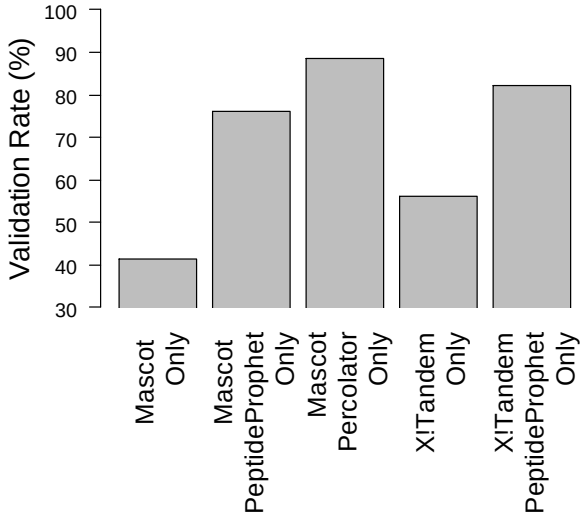

Figure 4.4 The validation rates of tool-specific PSMs.

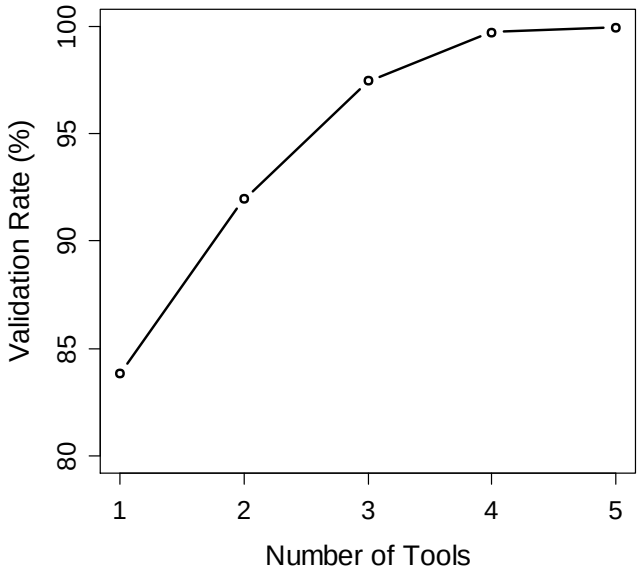

Figure 4.5 Validation rate as a function of number of tools by which PSMs can be identified.

Due to the fact that PSMs that are identified by multiple tools show high reliability, it was interesting to investigate if the best way to generate reliable results is by overlapping results from different statistical tools. Therefore, the overlapped PSMs between two or more statistical tools were investigated, namely the combination of Mascot and Mascot PeptideProphet, Mascot and Mascot Percolator, Mascot and X!Tandem, Mascot PeptideProphet and Mascot Percolator, Mascot PeptideProphet and X!Tandem PeptideProphet, X!Tandem and X!Tandem PeptideProphet, and Mascot, Mascot PeptideProphet and Mascot Percolator. As summarized in Table 4.4, the combinations of tools had validation rates no less than 99.4%. Compared to the overall validation rate of PSMs, which is 97.1%, higher reliability is

shown. Of those combinations, Mascot PeptideProphet and Mascot Percolator provided the highest number of validated PSMs while maintaining a similarly high validation rate (99.4%). However, there was a downside to this strategy as well. Overlapping results from different statistical tools decreased the total number of PSMs. Compared to the number of KR PSMs (119,105) identified by Mascot PeptideProphet alone, the combination of Mascot Percolator and Mascot PeptideProphet only had 112,365, a noticeable 5.7% drop in the number of total PSMs observed. In order to achieve the same validation rate by adjusting the probability cut-off in Mascot PepetideProphet, a bigger drop (12.6%) in the total PSM numbers was observed. Clearly, even though both methods could improve the reliability of consequent PSMs, simply overlapping results from different statistical tools could achieve the same reliability without suffering as much loss in identification number.

Table 4.4 Combination of Statistical Tools

Tools	Mascot & Mascot PeptideProphet	Mascot & Mascot Percolator	Mascot & X!Tandem	Mascot PeptideProphet & Mascot Percolator	Mascot PeptideProphet & X!Tandem PeptideProphet	X!Tandem & X!Tandem PeptideProphet	Mascot & Mascot PeptideProphet & Mascot Percolator
KR PSMs	93,186	91,213	78,734	112,365	98,076	96,873	91,098
Validated KR PSMs	92,900	91,052	78,710	111,675	97,904	96,319	90,946
Validation Rate	99.7%	99.8%	100.0%	99.4%	99.8%	99.4%	99.8%

A more detailed examination was performed by analyzing those statistical tools individually. First, the original Mascot result was examined. As shown in Table 4.3, Mascot has the highest validation rate (99.6%) among all the tools used in this study. It demonstrated the great specificity of Mascot's original scoring algorithm. At the same time, with the lowest number of validated KR PSMs (93,041, 68.4% of the total validated KR PSMs) inferior sensitivity compared to other statistical tools used in this study is suggested. As discussed in the result summary session, one widely believed suspect of the poor sensitivity is the high identity threshold assigned by the Mascot algorithm. It is very likely that a PSM has a decent Mascot ion score but is still considered insignificant because it fails to pass the identity threshold. Therefore, the relationship between the original Mascot score and the chance of the corresponding PSMs being validated was inspected. As illustrated in Figure 4.6, a trend was observed: as the original Mascot ion score increased, it was more likely that the corresponding PSMs were validated. When PSMs had Mascot ion scores lower than 20, the validation rate was only 86.3%. When PSMs had Mascot ion scores higher than 35, almost all the PSMs (99.9%) could be validated. It is intuitive that a higher Mascot ion score corresponds to a lower likelihood of the PSM being a random assignment. Interestingly, when PSMs have ion scores between 25 and 30, most of them are considered insignificant by Mascot as they failed to pass their identity thresholds. However, they still demonstrated an acceptable validation rate (94.8%). In fact, if the identity threshold was set to be 25, the overall validation rate would only decrease slightly from 99.6% to 99.0%. However, the number of validated PSMs

would increase by 18.6% (17,453 validated PSMs). Using the validation method, it was experimentally confirmed that the origin of the poor sensitivity comes from an unreasonably high identity threshold of Mascot. To address this issue, one might consider using different statistical strategies to evaluate the significance of PSMs (e.g., target-decoy strategy, Mascot Percolator or Mascot PeptideProphet).

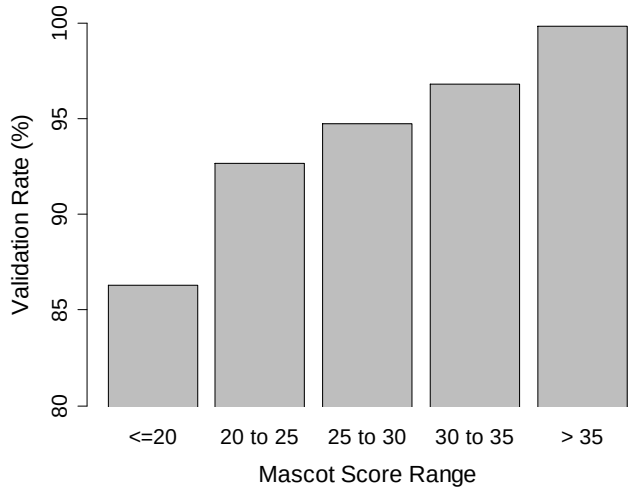

Figure 4.6 The validation rates of PSMs within different Mascot score ranges.

Of those strategies, the target-decoy strategy is one of the most popular and straightforward to use. It is often considered to be the gold standard of global FDR estimation of peptide identifications. However, some researchers[29] have expressed concerns that since the source code of Mascot is not publicly available, it is difficult to conclude whether or not Mascot is compliant with the principles of the target-decoy strategy. In this work, a direct and unbiased evaluation was provided on how accurately this strategy can estimate the true error rate of Mascot results. As shown in Figure 4.7, the global FDR estimated by the target-decoy strategy was directly compared with the true error rate (1 - validation rate). If the

target-decoy strategy can provide a reliable and unbiased assessment of the true error rate of the Mascot result, a straight 45 degree line should be observed. In Figure 4.7, the global FDR of the Mascot result estimated by the target-decoy strategy strongly agreed with the experimental error rate. It indicated that such statistical estimation is an accurate reflection of the true error rate of the Mascot result. Moreover, if the global FDR was slightly increased from 0.4% (significance threshold = 0.05) to 0.8%, the number of true positives can be improved by 25.3% (23,577 validated PSMs). Clearly replacing the identity threshold with a global FDR threshold estimated by target-decoy strategy is one possible alternative to avoid poor sensitivity.

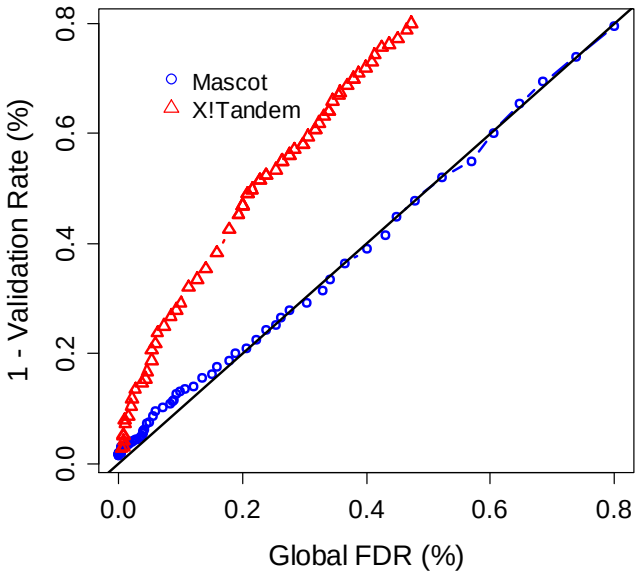

Figure 4.7 1 - validation rate as a function of estimated global FDR by the target-decoy approach.

On the contrary, the global FDR of the X!Tandem result estimated by the target-decoy strategy did not appear to agree with the experimental error rate as

well as Mascot did. It showed that such statistical evaluation underestimated the true error rate of the X!Tandem result. This underestimation of global FDR could be ascribed to the multi-stage function (refinement option) of X!Tandem searches. In X!Tandem searches, the refinement option allows X!Tandem to first filter the concatenated database before exploring every conceivable identification, which increases the proportion of target sequences in the remaining database and consequently biases against the decoy search. As a result, the number of decoy matches will not be a truthful reflection of incorrect matches in the target search, giving rise to unfair underestimation of global FDR for X!Tandem results. This compliance issue has been brought up and widely discussed in several studies[30-32]. Now using the validation method, it has been experimentally confirmed. Therefore it is reasonable to state that target-decoy strategy might not be the best choice to estimate global FDR with respect to multi-stage search engines like X!Tandem.

Finally, it is of great interest to know which statistical tool performs the best. Since every tool possesses its own way to assess the reliability of the result, it is difficult to compare them on the same scale. However, thanks to the advantage of experimental validation, not only can one calculate the true error rate of results from each tool, it is also possible to put them on the same scale to do the comparison. The number of validated PSMs was plotted against true error rate (1 – validation rate) for every statistical tool to investigate how well those tools perform at different error levels. In Figure 4.8, the number of validated PSMs truthfully reflects the number of true positives of each search result, while 1-

validation rate represents the true error rate of each search result. As shown in Figure 4.8, Mascot Percolator appeared to outperform the other four statistical tools with respect to sensitivity and specificity at all error levels. Among Mascot related results, the identification number improvement from both Percolator and PeptideProphet still held true, suggesting two possible alternatives to alleviate the unduly conservative problem of Mascot. The identification improvement of X!Tandem from PeptideProphet was confirmed as well. Therefore, it is reasonable to conclude that using an advanced statistical tool to process the original result further is certainly advantageous. However, it is worth mentioning that all the comparisons were done at a low error rate range. Because the validation is not applied to all the spectra but only the PSMs identified by those statistical tools at different thresholds, it is not feasible to extend the comparison to a higher error rate range. At higher error rate regions more studies are needed before any conclusions can be drawn.

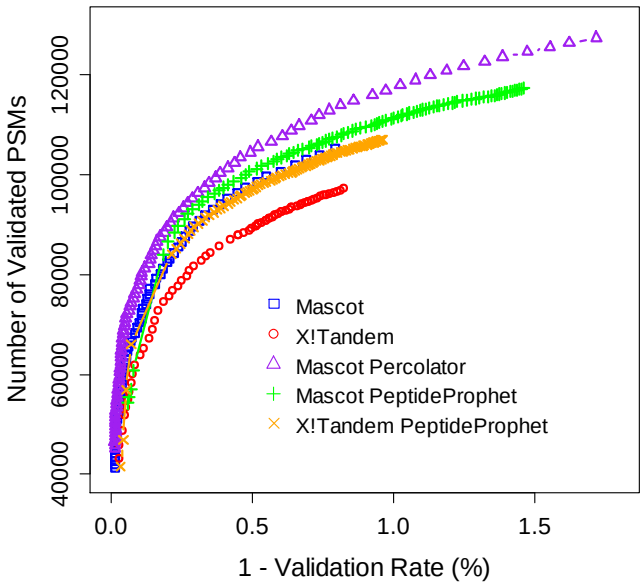

Figure 4.8 Number of validated PSMs as a function of 1 - validation rate.

4.4 Conclusions

In this work, an inclusion strategy and ^{18}O-labeling method were developed to experimentally validate PSMs. Using the experimental validation method, isolation of true PSMs from all the PSMs becomes possible. With those validated PSMs, the performance of commonly used search engines (Mascot and X!Tandem) and two popular statistical approaches (PeptideProphet and Percolator) were carefully examined. In this study, it was found that PSMs identified by multiple tools had lower error rates than the ones identified by only one tool. By comparing the numbers of validated PSMs at the same validation rate, it was found that it was advantageous (more validated PSMs) to embrace the overlapped PSMs from multiple statistical tools than simply to raise the score threshold. Next, it was confirmed that the unreasonably rigorous identity threshold was the cause

of the poor sensitivity of Mascot when searching large space. Apart from using Mascot Percolator or Mascot PeptideProphet, using global FDR estimated by target-decoy strategy as a threshold instead of using Mascot identity threshold was a possible way to improve its sensitivity. Besides, it was also confirmed that X!Tandem (with refinement function on) is not compliant with target-decoy strategy. Moreover, it was found that applying Percolator or PeptideProphet to original search results could truly improve the number of true PSMs while maintaining a relatively low error rate. Finally, the investigation on the performance of all five statistical tools revealed that Mascot Percolator outperformed the other four statistical tools.

This study was focused on the development of an experimental approach to validate PSMs and using those validated PSMs to examine the performance of statistical tools on the peptide level. In the near future, a detailed study at the protein level will be carried out using the same validated dataset. Moreover, after the experimental validation process, all the validated unlabeled PSMs are ready to be compiled, processed and stored into a spectral library for future usage. The detail procedure of constructing a spectral library can be found in the experimental section of Chapter 3.

4.5 Literature Cited

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198-207.

(2) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods* **2007**, *4*, 787-797.

(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(4) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426-1436.

(5) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(6) Deutsch, E. W.; Lam, H.; Aebersold, R. *Physiol. Genomics* **2008**, *33*, 18-25.

(7) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. *Mol. Cell. Proteomics* **2008**, *7*, 962-970.

(8) Wang, N.; Li, L. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1573-1587.

(9) Fenyoe, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.

(10) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(11) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(12) Kaell, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

(13) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.

(14) Qian, W.-J.; Liu, T.; Petyuk Vladislav, A.; Gritsenko Marina, A.; Petritis Brianne, O.; Polpitiya Ashoka, D.; Kaushal, A.; Xiao, W.; Finnerty Celeste, C.; Jeschke Marc, G.; Jaitly, N.; Monroe Matthew, E.; Moore Ronald, J.; Moldawer Lyle, L.; Davis Ronald, W.; Tompkins Ronald, G.; Herndon David, N.; Camp David, G.; Smith Richard, D. *J. Proteome Res.* **2009**, *8*, 290-299.

(15) Miyagi, M.; Rao, K. C. S. *Mass Spectrom. Rev.* **2007**, *26*, 121-136.

(16) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015-1024.

(17) Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002**, *20*, 163-170.

(18) Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; Von Haller, P.; Aebersold, R.; Kolker, E. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1214-1221.

(19) Pratt, J. M.; Robertson, D. H. L.; Gaskell, S. J.; Riba-Garcia, I.; Hubbard, S. J.; Sidhu, K.; Oliver, S. G.; Butler, P.; Hayes, A.; Petty, J.; Beynon, R. J. *Proteomics* **2002**, *2*, 157-163.

(20) Volchenboum, S. L.; Kristjansdottir, K.; Wolfgeher, D.; Kron, S. J. *Mol. Cell. Proteomics* **2009**, *8*, 2011-2022.

(21) Xu, M.; Li, L. *J. Proteome Res.* **2011**, *10*, 3632-3641.

(22) Wu, F.; Wang, P.; Zhang, J.; Young, L. C.; Lai, R.; Li, L. *Mol. Cell. Proteomics* **2010**, *9*, 1616-1632.

(23) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. *Nat. Biotechnol.* **2006**, *24*, 1285-1292.

(24) Tabb, D. L. *J. Proteome Res.* **2008**, *7*, 45-46.

(25) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646-4658.

(26) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 286-292.

(27) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 254-265.

(28) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. *Proteomics* **2010**, *10*, 1150-1159.

(29) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111-1120.

(30) Everett, L. J.; Bierl, C.; Master, S. R. *J. Proteome Res.* **2010**, *9*, 700-707.

(31) Tharakan, R.; Edwards, N.; Graham, D. R. M. *Proteomics* **2010**, *10*, 1160-1171.

(32) Bern, M.; Phinney, B. S.; Goldberg, D. *J. Proteome Res.* **2009**, *8*, 4328-4332.

Chapter 5

Strategies for Identification of Single-hit Proteins with High Confidence*

5.1 Introduction

Throughout the past decade, several sophisticated database search engines, such as Mascot[1], SEQUEST[2] and X!Tandem[3], have been developed to correlate tandem mass spectra (MS/MS) with peptide sequences and consequently infer protein identification. This widely used strategy is called bottom-up proteomics[4]. While these search engines have been proven to be robust in proteomic studies, their results are not error-free. The resultant peptide-spectrum matches (PSMs) often require statistical assessments, either individual or global, to attain a final identification list that is deemed to be correct identifications at a defined confidence level[5]. For instance, a probability-based Mascot ion score and identity threshold are implemented in the Mascot algorithm to measure reliability of each individual PSM (www.matrixscience.com). Alternatively X!Tandem arrives at the same goal by reporting expectation values (E-values) of PSMs[6]. In addition, estimation of the global false discovery rate (FDR) of peptide identifications can be achieved by using the target-decoy approach. Moreover, sophisticated statistical tools, such as Percolator[7, 8] and PeptideProphet[9] have been created to

* A version of this chapter has been prepared for submission as Xu, M, Li, L., Strategies for Identification of Single-hit Proteins with High Confidence.

provide both a global and local FDR estimation of peptide identifications. At the protein level, several tools, such as ProteinProphet[10], Mayu[11] and MS-GF[12], have also been developed to estimate global and local FDRs.

Even though so many attempts have been successfully made to gauge the reliability of peptide and protein identifications, many researchers are still cautious about reporting protein identifications[13-15]. In principle, a protein is identified when at least one of its associated peptides is matched (assuming the same peptide sequence is not present elsewhere in the proteome). However, because of the error-prone property of peptide identification tools, it has become common practice to selectively report proteins with at least two unique peptides (multi-hits) as reliable identifications while ignoring proteins with only one unique peptide (single-hits) if no additional corroboration can be provided. Theoretically, this "two-peptide rule" is based on the multiplication rule of independent events in probability theories. If a protein is identified by two different peptide sequences, the probability of this protein being a random match is equal to the multiplication of probability of each peptide being random. Intuitively, this rule can provide a stringent control on the quality of protein identifications. In practice, however, this rule appears to be unduly conservative and causes enormous protein information loss (approximately one third of all protein identifications[16]). In a recent study[12], Gupta and Pevzner argued that the "two-peptide rule" should be abandoned and single-hits should be treated at par as multi-hits with control of the protein FDR. While we fully agree with the argument that the commonly used "two-peptide rule" jeopardizes the sensitivity-

specificity trade-off in protein identifications, in our opinion more studies, especially experimental validation studies, should be carried out to evaluate the accuracy of protein level FDR estimation before "the two-peptide rule" is completely replaced with protein level FDR gauges.

In Chapter 3 and 4, two approaches have been successfully developed to experimentally validate PSMs. By comparing unlabeled and isotope-labeled matches of the same peptide sequence, it was demonstrated that not only is experimental validation a good way to isolate correct PSMs from search results and thus construct spectral libraries, but it can also be used to assess how accurate statistical tools estimate error rate of search results at the peptide level[17, 18]. In the experimental workflows of Chapter 4, an inclusion strategy was used to perform targeted analyses for ^{18}O-labeled PSMs in the LC-MS/MS runs. With multiple inclusion runs, it was possible to find almost all ^{18}O-labeled counterparts of pre-identified unlabeled PSMs. Using a retention time-based data filtering strategy, true and false identifications in the unlabeled PSMs were easily distinguished. Based on those experimentally validated PSMs, the performance of some commonly used statistical tools, such as Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet, was carefully inspected.

Since the experimental validation approach enabled the differentiation between true and false peptide identifications, one can easily determine whether a protein identification is true or false by looking for a true peptide identification within. Due to the categorization of true and false protein identifications, protein

level FDRs can be truthfully calculated. In this work, a large experimentally validated human data set was used (over 400,000 spectra and 140,000 PSMs) to study the validity of protein identifications and consequently develop practical strategies to deal with single-hits. It was demonstrated that a simplistic peptide level FDR gauge cannot provide an acceptable error rate control for protein identifications. By further categorizing single-hits, a highly reliable subgroup can be isolated. Using multiple search engine results of the same data set or applying different thresholds on single-hits and multi-hits are good ways to report protein identifications.

5.2 Experimental

5.2.1 Data Sets

The MS/MS data set used in this study was obtained from human cell samples (SU-DH-L1 cells, a human lymphoma cell line[19]). The detailed protocol used to generate this dataset can be found in Chapter 4. Briefly, proteins were extracted from human cells and then fractionated by reversed phase liquid chromatography. Next, each fraction of proteins was digested into peptide mixtures by trypsin. The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA) to collect MS/MS spectra. In total, 401,762 MS/MS spectra were collected.

5.2.2 Database Search

Using Proteinlynx Global Server 2.3.0 (Waters) all raw LC-ESI data were lock-mass corrected, de-isotoped, and converted to peak list files. A concatenated database was constructed by combining the IPI human database (version 3.68, size ≈ 48 MB, 87,061 sequences) and its reserved proteome sequences together. Database searches were carried out using both Mascot (version 2.2.1) and X!Tandem (The Global Proteome Machine Organization, 2007.07.01).

In the Mascot search, the search parameters were selected as follows: enzyme, trypsin; missed cleavages, 1; peptide tolerance, 30 ppm; MS/MS tolerance, 0.2 Da; fixed modification, carbamidomethyl (C); variable modifications, ammonia-loss (N-term C), N-Acetyl (protein), oxidation (M), pyro-Glu (N-term Q), and pyro-Glu (N-term E).

In the X!Tandem search, the search parameters, including the variable modification settings in the refinement function, were kept the same as Mascot searches. The search was restricted to the aforementioned concatenated database.

5.2.3 Peptide Identification

In this study, five different statistical tools were used to identify peptides from MS/MS spectra, including the original Mascot, the original X!Tandem, Mascot Percolator, Mascot PeptideProphet and X!Tandem PeptideProphet.

In the original Mascot search result, the significance threshold of 0.05 was applied to all PSMs. Therefore, if a PSM has a Mascot ion score no less than its Mascot identity threshold, it was deemed as identified. By the definition of

Mascot scoring algorithm, the application of the significance threshold of 0.05 ensures that the probability of an identified PSM being random in the identification list is no more than 5%.

X!Tandem implements a different scoring scheme compared to Mascot. In X!Tandem E-value (Expectation value) is calculated for each PSM to provide statistical evaluation. In its definition[6], E-value is the number (not probability) of random matches that would be expected to have the same or better scores. Therefore the higher the E-value, the less likely such PSM is to be deemed as a valid identification. Here, in the original X!Tandem search result, an arbitrary maximum E-value for PSM is set to 0.05.

Mascot Percolator[7] was also used to statistically evaluate the Mascot result so as to improve the number of identifications. After being processed by Percolator, each PSM is assigned with two statistical values, posterior error probability (PEP) and q-value. The q-value of each PSM can be understood to be the minimal global FDR required to include such PSM in the search result[8, 20]. PEP can be deemed as the local FDR of a PSM, which indicates the probability of such match being random. In Mascot Percolator processed results, the original Mascot p-values are replaced with PEP values. As a result based on the Mascot scoring equation, Mascot Score = $-10 \times \log_{10}(p)$, the new Mascot score of 13 represents that the local FDR of such PSM is 0.05. By applying the Mascot score threshold of 13 to the Mascot Percolator result, it is guaranteed that the maximum

local FDR of the search result is 5%. In this study, the minimum new Mascot score of 13 was chosen.

PeptideProphet[10, 21] was the other machine learning algorithm used to statistically evaluate the results both from Mascot and X!Tandem. As a part of the statistical software called Trans Proteomic Pipeline (version 4.3, JETSTREAM REV 1, Build 200909091257 MinGW), PeptideProphet assigns a probability to each PSM. Based on the definition of PeptideProphet probability, it is the probability of a PSM being correct. Therefore, 1 – probability can be understood as the local FDR of a PSM. In this study, a concatenated search and non-parametric modeling[22-24] were used. The negative distribution can be readily pinned down by decoy hits when fitting the bimodal distribution in PeptideProphet to maximize peptide identification while maintaining low error rate. In PeptideProphet treated results, the minimum probability of 0.95 (local FDR of 0.05) was chosen.

5.2.4 PSM Validation

In Chapter 4, a strategy was devised to experimentally validate identified PSMs. The concept of this experimental validation is fairly straightforward. First, the ^{18}O-labeling protocol was applied to each fraction of peptide digests to label all the tryptic peptides with C-terminal lysine or arginine (KR PSMs). Next, based on the masses of all the unlabeled KR PSMs, mass-to-charge ratios of their ^{18}O-labeled counterparts can be readily calculated. Combined with retention information, inclusion lists for the ^{18}O-labeled RPLC-MS/MS runs can be easily

generated. According to the inclusion list, multiple targeted RPLC-MS/MS analyses were performed on all the ^{18}O-labeled fractions in order to identify the ^{18}O-labeled counterparts of all the unlabeled KR PSMs. Finally, a comparison was carried out between the eluting organic composition (%B) of the unlabeled and labeled identifications of the same peptide sequence. If they differed by more than 0.5%, the unlabeled PSM was not to be considered validated by that labeled PSM even if they appear to share the same sequence. Theoretically, if an unlabeled KR PSM is a correct identification, its ^{18}O-labeled counterpart should be also be identified with similar %B. After applying the validation method, all the unlabeled KR PSMs can be categorized into two groups, validated KR PSMs and invalidated KR PSMs. The validated KR PSMs can be deemed as true positives, while the invalidated ones are false positives.

5.2.5 Protein Identification

Due to the limitation of the validation method, only KR PSMs were used to infer protein identification. A protein that has at least one peptide score above the chosen threshold is considered to be identified. A protein that has at least one validated peptide score above the chosen threshold is considered validated. According to the number of unique peptides a protein has, protein identification can be categorized into two groups, single-hit (only one unique peptide) and multi-hit (more than one unique peptide). Furthermore, the name homologous single-hit (HSH) is coined to describe the single-hits that share at least one common peptide with other protein identification(s). Alternatively, strict single-

hit (SSH) is used to describe the single-hits that only have one unique peptide and share no common peptide with any other protein identification.

5.2.6 Data Processing

All in-house programs were written in Perl 5.12 (http://www.perl.org). Charts and graphs were generated using R's plotting packages (http://www.r-project.org/) and Microsoft Excel 2007. Software was run on standard desktop and laptop computers running Windows 7 (Home Edition).

5.3 Results and Discussion

From the RPLC-QTOF MS/MS analyses, a total of 401,762 spectra were collected. All the spectra were searched using both Mascot and X!Tandem search engines and all the potential PSMs were given scores or probabilities after being statistically evaluated by Mascot, PeptideProphet, Mascot Percolator and X!Tandem. Arbitrary score thresholds were set up for all the statistical tools. For original Mascot, the significance threshold of 0.05 was used. For PeptideProphet results, the minimal probability of 0.95 was applied. For Mascot Percolator results, the new Mascot Identity threshold of 13, which is equivalent to a maximum local FDR of 0.05, was applied. In original X!Tandem results, the maximum expect value of 0.05 was adopted.

As shown in Table 5.1, 94,121 PSMs were identified by original Mascot search engine with an estimated global FDR of 0.4%. Among those PSMs, 93,381 of them are matches with C-terminal lysine or arginine (KR PSMs). After evaluating Mascot results by PeptideProphet or Mascot Percolator, the total numbers of PSMs were evidently higher, 120,105 and 130,774, increasing the original Mascot result by 27.6% and 38.9%, respectively. At the same time, Mascot PeptideProphet maintained a relative low error rate of 0.5% while Mascot Percolator had a q-value score of 0.4%, indicating a similar error rate as the original Mascot result. When processing all the MS/MS spectra by X!Tandem search engine, 99,186 PSMs were collected, among which 98,175 can be categorized as KR PSMs. After processing X!Tandem results with PeptideProphet, 108,150 KR PSMs was observed. It was found that the global FDR was 0.4% in both X!Tandem and X!Tandem PeptideProphet results. Further analysis at peptide level shows that 77,924 PSMs can be found by all five statistical tools, constituting the majority of all the KR PSMs, 55.6% (Figure 5.1A). As expected, each statistical tool has its own strengths when matching peptide sequences to MS/MS spectra. As shown in Figure 5.1A, the original Mascot appears to be the most agreeable statistical tool, showing only 75 unique PSMs. Meanwhile, Mascot Percolator seems to be the most sensitive tool, contributing 92.5% of all the KR PSMs.

Due to the advantages of our ^{18}O-labeling validation approach, the correct PSMs were able to be differentiated from spurious matches. Briefly, if an unlabeled KR PSM is correct, its ^{18}O-labeled counterpart should be confidently

identified with similar %B as well. After applying the validation method, all the unlabeled KR PSMs can be categorized into two groups, validated KR PSMs and invalidated KR PSMs. The validated KR PSMs can be deemed as true positives, while the invalidated ones are false positives. Based on the number of true and false positives, PSM validation rates can readily calculated for results from all statistical tools. As shown in Table 5.1, in the original Mascot result, 93,041 out of 93,381 KR PSMs can be experimentally validated. The PSM validation rate is 99.6%. Similarly, the PSM validation rates for the other statistical tools are 98.5%, 98.3%, 99.2% and 99.0% for Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet, respectively.

Table 5.1 Identification and Validation Result Summary.

	Mascot	Mascot PeptideProphet	Mascot Percolator	X!Tandem	X!Tandem PeptideProphet
All the PSMs	94,121	120,105	130,774	99,186	109,229
KR PSMs[a]	93,381	119,105	129,634	98,175	108,150
Proteins[b]	3,256	4,114	3,933	3,476	3,607
Multi-hits[c]	2,246	2,588	2,730	2,392	2,507
Single-hits[d]	1,010	1,526	1,203	1,084	1,100
SSHs[e]	899	1,362	1,059	982	989
HSHs[f]	111	164	144	102	111
Validated KR PSMs	93,041	117,363	127,405	97,364	107,108
Validated Proteins	3,180	3,612	3,618	3,285	3,386
Validated Multi-hits	2,246	2,580	2,714	2,388	2,502
Validated Single-hits	934	1,032	904	897	884
Validated SSHs	823	870	761	795	773
Validated HSHs	111	162	143	102	111
PSM Validation Rate[g]	99.6%	98.5%	98.3%	99.2%	99.0%
Protein Validation Rate	97.7%	87.8%	92.0%	94.5%	93.9%
Multi-hits Validation Rate	100.0%	99.7%	99.4%	99.8%	99.8%
Single-hits Validation Rate	92.5%	67.6%	75.1%	82.7%	80.4%
SSHs Validation Rate	91.5%	63.9%	71.9%	81.0%	78.2%
HSHs Validation Rate	100.0%	98.8%	99.3%	100.0%	100.0%

a. KR PSMs: PSMs with C-terminal lysine or arginine;

b. Proteins: protein identifications inferred from KR PSMs;

c. Multi-hit: protein identifications with at least two unique peptides;

d. Single-hit: protein identifications with only one unique peptide;

e. SSHs (Strict Single-Hits): the single-hits that only have one unique peptide and share no common peptide with any other protein identification;

f. HSHs (Homologous Single-Hits): the single-hits that share at least one common peptide with other protein identification(s);

g. Validation Rate is calculated by dividing the number of validated matches (peptides or proteins) by the total number of matches. It can be understood as 1 – FDR.

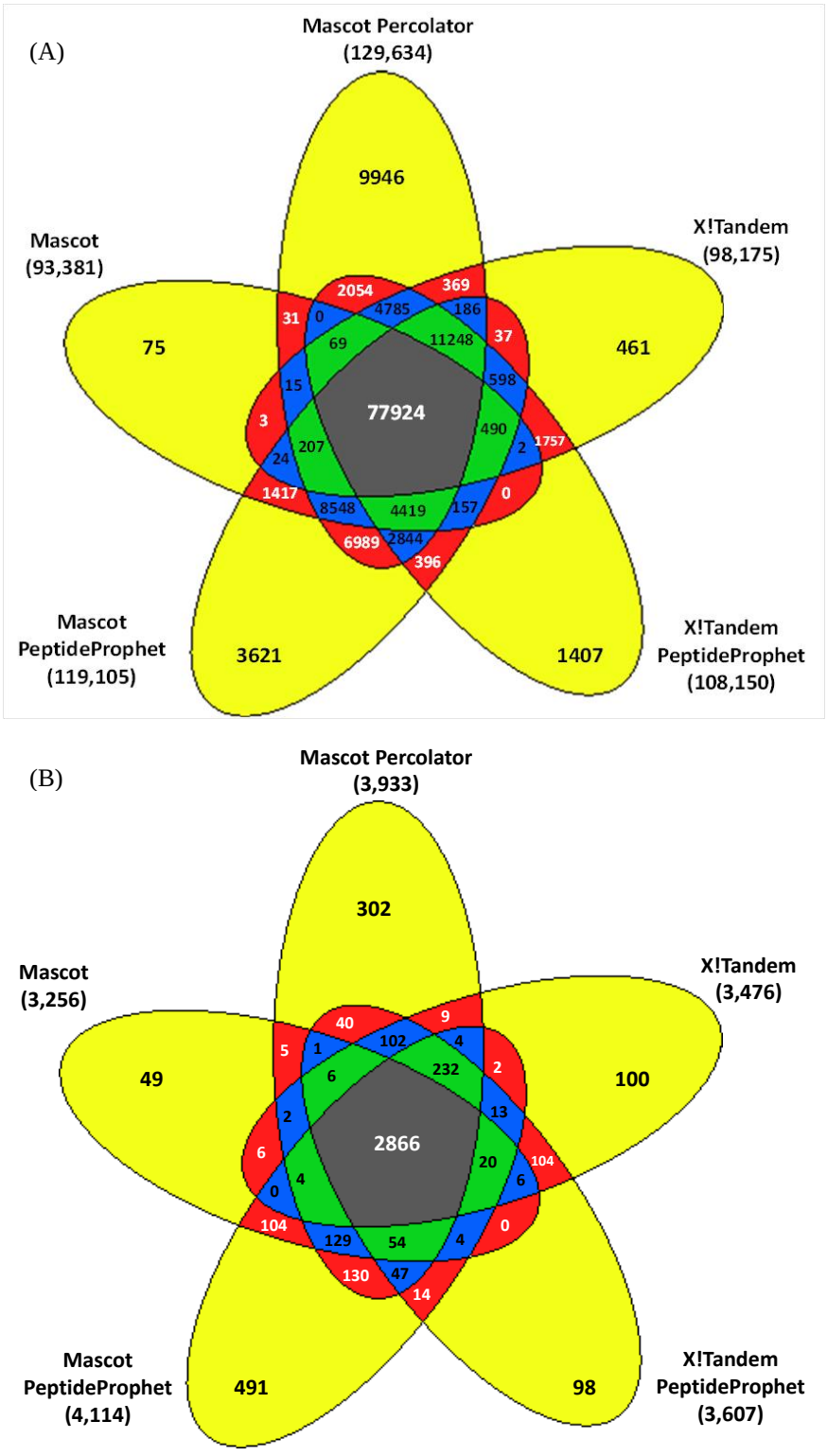

Figure 5.1 Venn diagram analysis of (A) all the KR PSMs and (B) protein identifications from Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet.

Based on the all the KR PSMs, the number of protein identifications were inferred for the result from each statistical tool. As shown in Table 5.1, the protein identification numbers are 3,256, 4,114, 3,933, 3,476 and 3,607 for Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet, respectively. Interestingly, at the peptide level, Mascot Percolator showed the highest identification number (129,634 PSMs). Meanwhile, Mascot PeptideProphet appeared to have the highest number of protein identifications. Clearly, maximized number of peptide identifications does not necessarily give rise to the highest number of protein identifications. It happens that peptides of several proteins are identified in one result, while peptides of various proteins are matched in another. It was also observed that different statistical tools matched different numbers of single-hit protein identifications, ranging from 1,010 to 1,527 (see Table 5.1). They consisted of approximately one third of the total number of protein identifications in the result from each statistical tool and were normally not considered as confident as multi-hits. Conventionally, researchers only report proteins with at least two peptides as confident identifications. Without additional corroboration, those single-hits eventually would end up being ignored, which inevitably leads to a substantial loss of information. However, including all single-hits might impair the reliability of the results. Therefore, how to report identification results of shotgun proteomic experiments is of great importance. With the assistance of our validated PSMs, better ways of reporting protein identifications may be found.

Before diving into the validation results, all the protein identifications from each tool were compared. As shown in Figure 5.1B, 2,866 proteins can be found by all five statistical tools, constituting 58.0% of all protein identifications. While the results from the statistical tools tested agreed well for most protein identifications, there were still some protein identifications unique to one statistical tool. Tool-specific protein identifications are coined to describe those proteins. From the previous study on peptide identifications (see Chapter 4), it was observed that the common identifications were more reliable than those tool-specific identifications. Overlapping identifications from different results might shed some light on how to process protein identifications and eventually generate a reliable report without losing too much information. Using validated results from Chapter 4, that question can surely be answered.

Since all the KR PSMs can be categorized into validated (true positives) and invalidated KR PSMs (false positives) after PSM validation process, it is not difficult to infer the validity of protein identifications. In principle, as long as the protein identification has one validated unique peptide identification, it is considered a validated protein identification (true positive). Based on the number of validated protein identifications, the protein validation rate can be easily calculated. As shown in Table 5.1, the majority of protein identifications can be validated. The protein validation rates were 97.5%, 87.5%, 91.8%, 93.8% and 93.6% for Mascot, Mascot PeptideProphet, Mascot Percolator, X!Tandem and X!Tandem PeptideProphet, respectively. Comparing to PSM validation rate of the same statistical tool, the corresponding protein validation rate was apparently

lower. This often occurs because random peptide matches (false positives) do not cluster to individual proteins at the same rate as correct peptide matches (true positives) do. Random peptide matches tend to cluster to single-hit proteins, while correct peptide matches are apt to cluster to multi-hit proteins. At the protein level, the number of false positives drops to much less than the number of true positive when inferring protein identifications from peptide matches. Understandably, a low global FDR at the peptide level (1 – PSM validation rate) does not necessarily result in a low global FDR at the protein level (1 – protein validation rate). For instance, in the Mascot PeptideProphet result, each PSM had a probability score of no less than 0.95, equivalent to a maximum local FDR of 0.05, and the global FDR at the peptide level of this result was as low as 1.5%, indicating high reliability. However, after protein inference, the global FDR at the protein level was found to be at an unacceptable level of 12.2%. Clearly, when the goal of the study is protein identification (e.g., biomarker discovery, proteome profiling), a simple FDR control at the peptide level is not sufficient to ensure the reliability of protein identifications, let alone the reliability of single-hits. With respect to single-hits, the number of validated proteins, validated single-hits and validated multi-hits and their validation rates were tabulated. Compared with the multi-hit validation rates, single-hit validation rates are constantly lower in all cases, indicating their inferior reliability. In fact, a student t-test was performed and the result showed that at the 99% confidence level the single-hit validation rate is significantly lower than the multi-hits validation rate. Considering that the average validation rate of multi-hits was 99.7%, it is evident that the single-hits

were the main source of false positives. In the case of the original Mascot result, in which multi-hits validation rate was as high as 100.0%, the single-hits were the only contributor of false protein identifications. Even with a low global FDR at the protein level (2.3% in the Mascot result), the reliability of single-hits still cannot be fully guaranteed (single-hits validation rate is 92.5% in Mascot).

From validation rates comparison, it is established that single-hits are not as reliable as multi-hits. Consequently, accepting all single-hits without any additional corroboration is not recommended. However, it does not mean that one should abandon single-hits all together. After all, the loss of one third of protein identifications is too costly. In order to resolve this dilemma, we first turned to the definition of single-hits. By definition[15], single-hits are the proteins identified by only one unique peptide. It was found that single-hits can be further categorized into two subgroups. The name homologous single-hit (HSH) was coined to describe the single-hits that share at least one common peptide with other protein identification(s), while strict single-hit (SSH) was used to describe the single-hits that only have one unique peptide and share no common peptide(s) with any other protein identification. Based on this categorization, it was found that HSHs only constituted on average 10.6% of the single-hits in each result. Nevertheless HSHs showed a significantly higher validation rate than SSHs. As shown in Table 5.1 and Figure 5.2, the average validation rate of HSHs is as high as 99.6%, comparable to the validation rate of multi-hits (99.7%). If multi-hits can be deemed as reliable identifications, so should HSHs. Consequently by simply further categorizing single-hits, homologous single-hits was isolated as a sub-

group with highly reliable identifications. It is reasonable to accept HSHs as reliable identifications as they have validation rate as high as multi-hits.

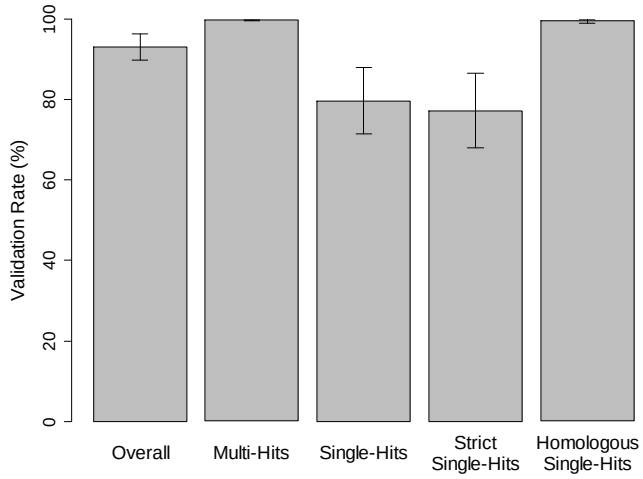

Figure 5.2 Validation rates for different types of protein identifications.

However, among all the single-hits, HSHs were not the major subgroup. SSHs consisted of 90.4% of all the single-hits on average. They still showed a relatively low validation rate (77.3% ± 9.2%), suggesting low reliability. Therefore, how to discern the true SSHs from false SSHs holds the key to improve the reliability of single-hits. Firstly, SSHs from different results were overlapped and analyzed. At the peptide level, common PSMs were found to be more reliable than those tool-specific identifications (see Chapter 4). At the protein level, a similar trend for SSHs should be discovered as well considering that a SSH only contains one unique PSM. Therefore, SSHs were categorized according to by how many tools they were identified and the validation rate in each category was calculated (see Figure 5.3). As expected, the majority of validated SSHs (88.5%) can be identified by more than one tool, while the

majority of invalidated SSHs (75.0%) can only be identified by one tool. The validation rate increased as the number of tools by which SSHs can be identified increased. The SSHs that were identified by at least 4 tools, had validation rates as high as 99.4% and constituted 67.5% of all the validated SSHs. Clearly, the number of tools by which SSH can be identified is a good indicator of whether such SSH is true or false identification.

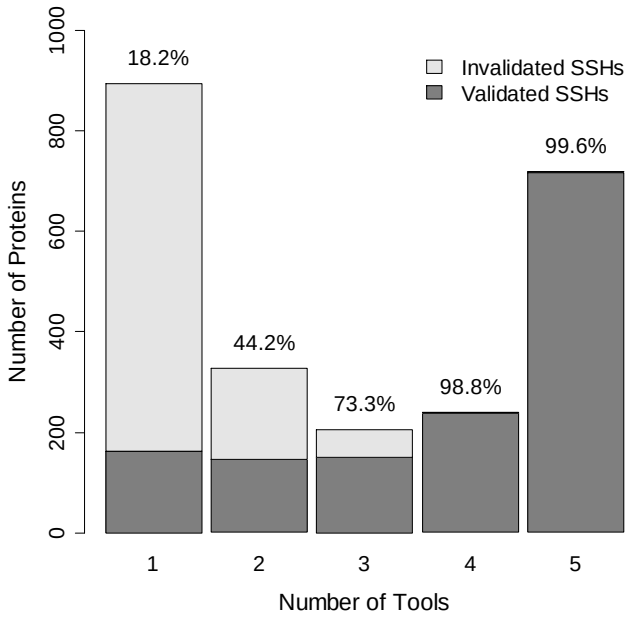

Figure 5.3 The numbers of SSHs (validated and invalidated) as a function of the number of tools by which SSHs can be identified. The percentage labels in the figure indicate the validation rates of SSHs.

In practice, not every research group has access to so many statistical tools. It is impractical to process single-hits that way. Since Mascot and X!Tandem are two of the most popular search engines, the common protein identifications from those two original results were first investigated. After overlapping all the protein identifications from both results, only 477 common SSHs can be found and their

validation rate was 99.2%. However, once the comparison was broadened from just among SSHs to all the protein identifications, allowing SSHs to compare with HSHs and multi-hits, it was found that 654 SSHs in the original Mascot result can also be identified by X!Tandem. The validation rate for those SSHs is 99.4%, indicating reliability as high as multi-hits (see Figure 5.4). Comparing to the total number of validated SSHs from Mascot result (823), simply overlapping protein identifications with the X!Tandem result managed to recover 79.5% of them. Overall, 3,011 proteins were identified with a global protein FDR of 0.1% (Table 5.2). Similarly, it is found that 555 SSHs in the original X!Tandem result can also be identified by Mascot with a validation rate of 99.3% (Figure 5.4). This simple overlapping method was able to salvage 69.2% of all the SSHs in X!Tandem. It appears that X!Tandem did a better job recovering SSHs in Mascot result than Mascot recovering SSHs in X!Tandem. This is due to the fact that there were fewer proteins identified in the original Mascot result. When comparing X!Tandem's SSHs with all the protein identifications in the Mascot PeptidePropeht result, which contained the highest number of protein identifications, 720 out of 795 validated SSHs (90.6% recovery rate) in X!Tandem were recovered. Those SSHs from X!Tandem result had a validation rate of 98.4%, still indicating high reliability. Overall, 3,045 proteins were identified with a global protein FDR of 0.2% (see Table 5.2). In conclusion, simply overlapping SSHs from one result with all the protein identifications in another is a good way to differentiate true and false SSHs.

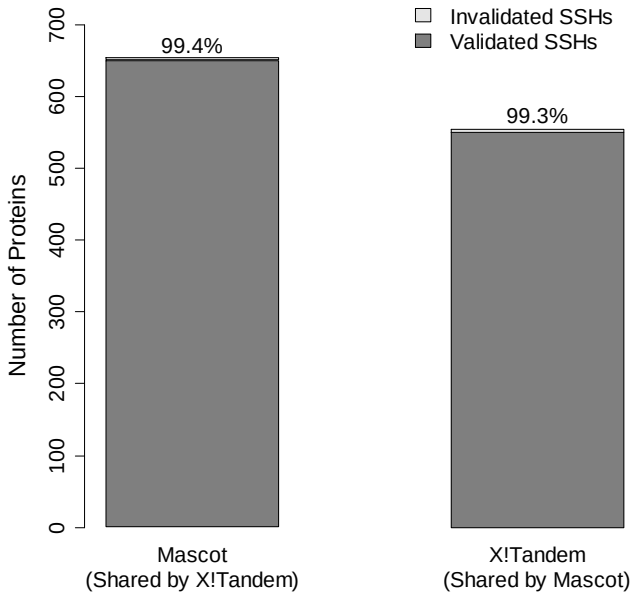

Figure 5.4 The number of SSHs (validated and invalidated) from Mascot that can be corroborated by X!Tandem and the The number of SSHs (validated and invalidated) from X!Tandem that can be corroborated by Mascot. The percentage labels in the figure indicate the validation rates of SSHs.

Overlapping results from two search engines is a simple and effective way to isolate true SSHs from false SSHs. Nonetheless it requires two different search engines. It might be inapplicable to a lot of researchers who only have access to one. Therefore if there was an approach that requires only the original software (e.g., Mascot or X!Tandem), it would be of great help to many researchers. First, the original Mascot result was examined. In principle, random peptide matches should generally have a lower Mascot ion score than correct ones. So by increasing the identity threshold of each PSM, the number of random matches would decrease.

Table 5.2 Comparison of Protein Report Approaches.

	Mascot			X!Tandem		
	Unique Peptides	Proteins	Global FDR (Protein)	Unique Peptides	Proteins	Global FDR (Protein)
Overlap[a]	16,447	3,011	0.1%	16,454	3,045	0.2%
Two-Stage[b]	16,306	2,870	0.5%	16,418	3,010	0.8%
One-Stage[c]	14,156	2,743	0.5%	13,320	2,907	0.8%

a. In the overlap approach, multi-hits and HSHs are accepted as confident protein identifications while only the SSHs inferred from the common PSMs found in both Mascot and X!Tandem results are reported.

b. In the two-stage approach, all the multi-hits and HSHs are inferred from PSMs that pass a relatively lenient score threshold (significance threshold of 0.05 in Mascot and maximum E-value of 0.05 in X!Tandem) while a more stringent threshold applies to PSMs that lead to SSHs (significance threshold of 0.01 in Mascot and maximum E-value of 0.005 in X!Tandem).

c. In the one-stage approach, all the PSMs used to infer protein identifications pass a stringent threshold (significance threshold of 0.01 in Mascot and maximum E-value of 0.005 in X!Tandem).

As shown in Figure 5.5A, as the identity threshold gradually increases, the validation rate of SSHs increases, indicating that the reliability of SSHs is increasing. However, this gain in reliability is not without any price. As illustrated in Figure 5.5A, as validation rate climbs up, the number of validated SSHs declines. Since the original identity threshold was obtained when the significance threshold of 0.05 was applied, based on the definition of Mascot identity threshold (Mascot identity threshold = $-10 \times \log_{10}(p/n)$), increasing the original identity threshold of each PSM by 7 would be equivalent to set the significance threshold to 0.01. Additionally, it was found that after increasing the identity threshold by 7, the validation rate of the remaining SSHs improved from

91.5% to 97.2%, indicating a less than 5.0% error rate among SSHs. The remaining SSHs constituted 60.6% of all the SSHs from the original Mascot result. In the original X!Tandem result a similar trend was found. As shown in Figure 5.5B, as the E-value cut-off decreases, indicated by the increasing values of –log(E) in the x-axis, the validation rate of SSHs increases and the number of validated SSHs declines. If a more stringent E-value cut-off (0.005) was chosen, 61.8% of validated SSHs can be recovered with a validation rate of 96.1%, indicating a less than 5.0% error rate among SSHs. It is therefore advantageous to utilize a two-stage scoring scheme to deal with protein identifications when only one search engine is available. In the first stage, a relatively lenient score cut-off (e.g., significance threshold of 0.05 in Mascot or maximum E-value of 0.05 in X!Tandem) is chosen for the all the PSMs. Next, all the peptide matches that are above the score cut-off are used to infer protein identifications. While keeping all the multi-hits and HSHs intact, applying a more stringent score cut-off (e.g., significance threshold of 0.01 in Mascot and maximum E-value of 0.005 in X!Tandem) on all the SSHs improves their reliability.

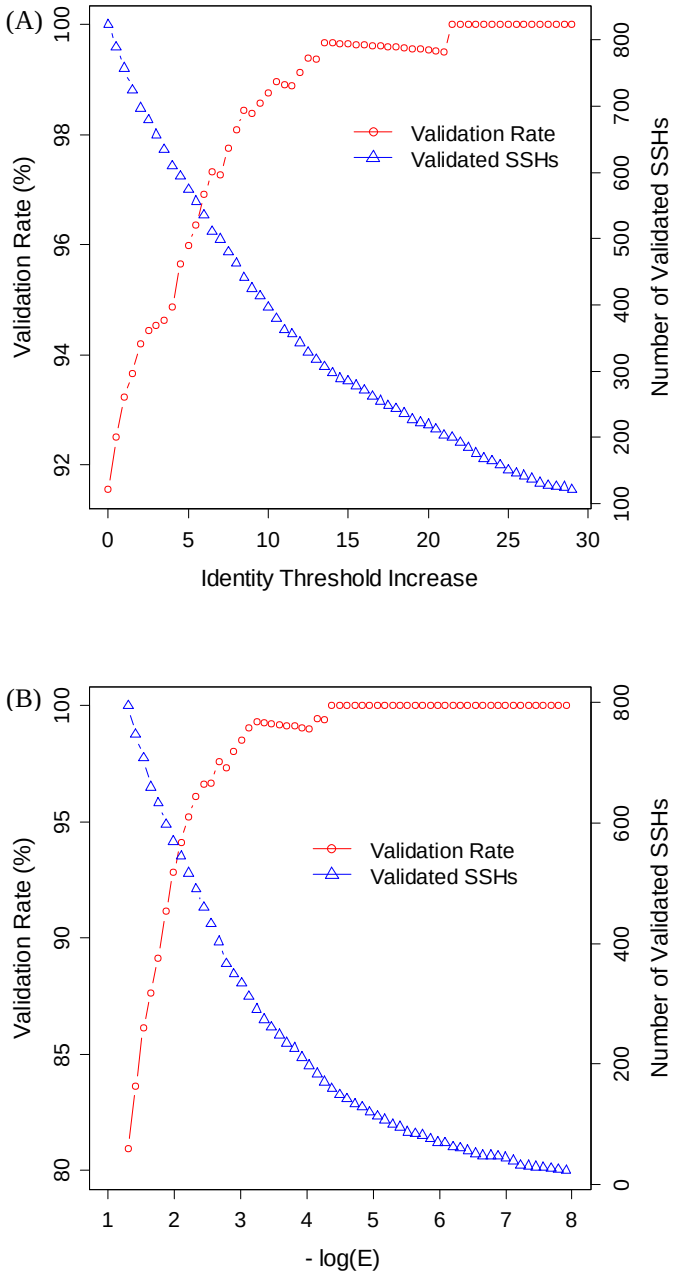

Figure 5.5 (A) The validation rate of SSHs and the number of validated SSHs as functions of the identity threshold increase in Mascot. (B) The validation rate of SSHs and the number of validated SSHs as functions of -log(E) in X!Tandem.

Since the application of a more rigorous score cut-off on SSHs can improve their collective reliability, should the same rigorous score cut-off be

adopted for all the PSMs? In our opinion the answer is no. Considering that the main source of errors in protein identifications lies in SSHs, the extra-stringent score cut-off on multi-hits or HSHs would not help reduce the error rate of the result. On the other hand, applying such stringent score cut-off on the entire result will definitely reduce the number of peptide identifications and consequently cause some information loss (e.g., reduced numbers of protein identifications and sequence coverage). The result of our two-stage method was compared with the result of applying a stringent score cut-off on all the PSMs (one-stage method). In Mascot, the two-stage method (applying significance threshold of 0.05 on PSM level and significance threshold of 0.01 on SSHs alone) outperformed the one-stage method (applying significance threshold of 0.01 on PSM level) with respect to the number of PSMs, unique peptide sequences and protein identifications. As shown in Table 5.2, the sole stringent identity threshold on all the PSMs can give rise to 13.2% and 4.4% reduction in unique peptides and protein identifications, respectively. At the same time, the global FDR on the protein level were the same for both methods, indicating the same high reliability of protein identifications. In X!Tandem, the superiority of the two-stage approach was also demonstrated.

5.4 Conclusions

In this study, experimentally validated PSMs from Chapter 4 were used to examine the validity of protein identifications from Mascot and X!Tandem results. The advantages of experimental validation made it became possible to isolate true

and false positives at both peptide and protein levels. With the numbers of true and false positives, one can readily calculate (not estimate) the true global FDRs of peptide and protein identifications. It was demonstrated that a low global FDR at the peptide level cannot guarantee a low global FDR at the protein level. If the goal of one's study is protein identification (e.g., biomarker discovery, proteome profiling), a simplistic global FDR control at the peptide level is insufficient to gauge the reliability of protein identifications. In this study, it was found that the commonly used "two-peptide rule" can in fact significantly improve the reliability of protein identifications but is unduly conservative. In order to retrieve the correct single-hits eliminated by the "two-peptide rule", a further categorization of all the single-hits discovered two subgroups: HSH and SSH. It was observed that HSHs were as reliable as multi-hits and thus recommended to be treated as such. With respect to the majority of single-hits, SSHs, two straightforward solutions were proposed to discern the true positives from false ones. If one has access to two search engines (e.g., Mascot and X!Tandem), one can compare SSHs from one result with all protein identifications from the other. The SSHs that can be found in both results are highly reliable and should be deemed as confident protein identifications. If there is only one search engine available, a two-stage threshold approach seems to be a rational choice. In the first stage, a relatively lenient score cut-off (e.g., significance threshold of 0.05 in Mascot or maximum E-value of 0.05 in X!Tandem) is chosen for the all the PSMs. Next, all peptide matches that are above the score cut-off can be used to infer protein identifications. While keeping all the multi-hits and HSHs intact, applying a more

stringent score cut-off (e.g., significance threshold of 0.01 in Mascot and maximum E-value of 0.005 in X!Tandem) on all the SSHs can improve their reliability. Using either approach, more protein identifications can be identified than the overly conservative "two-peptide rule" without sacrificing the global protein FDR.

5.5 Literature Cited

(1) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(2) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426-1436.

(3) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(4) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198-207.

(5) Nesvizhskii, A. I. *J. Proteomics* **2010**, *73*, 2092-2123.

(6) Fenyoe, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.

(7) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.

(8) Kaell, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

(9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(10) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646-4658.

(11) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. *Mol. Cell. Proteomics* **2009**, *8*, 2405-2417.

(12) Gupta, N.; Pevzner, P. A. *J. Proteome Res.* **2009**, *8*, 4173-4181.

(13) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y.-K.; Yoo, J.-S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M. *Proteomics* **2005**, *5*, 3226-3245.

(14) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. *Mol. Cell. Proteomics* **2006**, *5*, 787.

(15) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531-533.

(16) Wang, N.; MacKenzie, L.; De, S. A. G.; Zhong, H.; Goss, G.; Li, L. *J. Proteome Res.* **2007**, *6*, 263-272.

(17) Xu, M.; Li, L. *J. Proteome Res.* **2011**, *10*, 3632-3641.

(18) Zhong, H.; Marcus, S. L.; Li, L. *J. Proteome Res.* **2004**, *3*, 1155-1163.

(19) Wu, F.; Wang, P.; Zhang, J.; Young, L. C.; Lai, R.; Li, L. *Mol. Cell. Proteomics* **2010**, *9*, 1616-1632.

(20) Tabb, D. L. *J. Proteome Res.* **2008**, *7*, 45-46.

(21) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(22) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 286-292.

(23) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 254-265.

(24) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. *Proteomics* **2010**, *10*, 1150-1159.

Chapter 6

X!Tandem Percolator: Accurate and Sensitive Peptide Identification Tool*

6.1 Introduction

During the past decade tandem mass spectrometry has progressed to be a popular and powerful tool to study complex biological systems for proteomic studies[1-3]. In conjunction with liquid chromatography separations, thousands of tandem mass spectra are routinely acquired and in need of correlation to peptide sequences and eventually protein identifications. Instead of manual interpretation of each spectrum, search engines, such as Mascot[4] and X!Tandem[5], were developed to match the spectra with peptide sequences by comparing the experimental spectrum with the theoretical fragmentation patterns of individual peptide sequences derived from the protein sequences in a proteome database. To measure reliability of each individual peptide-spectrum match (PSM), a probability-based Mascot ion score and identity threshold are implemented in the algorithm of Mascot (www.matrixscience.com). When the significance threshold of 0.05 ($p = 0.05$) is applied, it ensures that the probability of an identified PSM being random in the identification list is no more than 5%. Meanwhile, X!Tandem

*A version of this chapter has been prepared for submission as Xu, M, Li, Z., Li, L., X!Tandem Percolator: Accurate and Sensitive Peptide Identification Tool. Zhendong Li contributed partially to data processing of this work.

accomplishes the same goal by reporting expectation values (E-values, the number of random matches that would be expected to have the same or better scores) of PSMs[6]. Afterwards, the concept of global false discovery rate (FDR) was proposed[7] as the standard to regulate the reporting of search results. The most common approach to estimate the global FDR of a search result is the target-decoy approach[8], which is based on the use of randomized decoy proteome databases. As the target-decoy approach enables the error control at the peptide level for different results from different search engines, it cannot provide any statistical evaluation on the reliability of each individual PSM.

In addition to the target-decoy strategy, more sophisticated algorithms were developed to provide both global FDR estimation as well as individual assessment of PSMs. They re-evaluate the qualities of PSMs from the original search result and assign new probability to each PSM by examining the properties of correct and incorrect PSMs. For instance, PeptideProphet[9] uses an expectation-maximization algorithm to fit the bimodal distribution formed by discriminant scores of correct and incorrect PSMs in the histogram and thus computes the probability of each PSM and global FDR of the entire result. Alternatively, Percolator[10] implements a different machine learning approach. After searching all the spectra in both target and decoy databases, Percolator extracts a vector of features that are related to the quality of the match (e.g., mass error and PSM score) from both target and decoy PSMs. Assuming that the features of correct matches (represented by high scoring target matches) differ from the features of incorrect matches (represented by decoy matches), an iterative classification

process is applied to find the best separation between correct and incorrect matches. After several iterations, the system converges and generates a robust classifier that can be used to calculate the probability of each PSM being random.

Among all the statistical strategies, Percolator was demonstrated to be one of the most sensitive and accurate tools to evaluate PSMs. Due to the adaptive nature of Percolator, Percolator has been successfully extended from the application of SEQUEST[11] results to the use of Mascot results[12]. When implementing Percolator with Mascot, the selection of the features used by Percolator was shown to be vitally important to the performance of Percolator. When the authors included extra features that included information such as intensity and fragment error, the sensitivity was boosted by 17%.

In this work, Percolator program has been successfully interfaced with X!Tandem using a very simple PHP program. Since it is critical to select the best discriminating features for Percolator as to achieve the best performance, a set of experimentally validated PSMs were used to optimize and validate our choice of features. In a previous study[13], Xu and Li described a method of using ^{15}N-labeling for validating the spectrum-to-sequence assignments and constructing a more reliable MS/MS spectral library (see Chapter 3). Due to the advantages of this experimental validation method, not only are a large set of spectrum-to-sequence assignments justified, the annotation of the spectra are also validated, giving us an opportunity to examine the spectral characteristics of true peptide identifications. By comparing the features of those experimentally validated

PSMs (34,993 MS/MS spectra) with those of false identifications, a comprehensive set of features can be chosen for Percolator in an objective and rational manner. Followed by the optimization of features, the accuracy of X!Tandem Percolator was demonstrated by comparing the estimated false discovery rate of the validated data set with the factual false discovery rate. By comparing the results from our X!Tandem Percolator and the original X!Tandem, superior sensitivity and specificity of the X!Tandem Percolator result was demonstrated. Lastly, X!Tandem Percolator was applied to results with various search conditions, such as large MS/MS data sets from different species, human (46,494 MS/MS spectra) and *E. coli* (88,306 MS/MS spectra) to examine robustness. As a result, our X!Tandem Percolator clearly improved the number of peptide identifications at the same level of FDRs in both cases.

6.2 Methods

6.2.1 Sample Preparation

Three different MS/MS data sets are used in this study. They are all originally collected using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). The experimental details can be found in the experimental section of Chapter 3 and 4.

6.2.2 *E. coli* Data Set

The *E. coli* K12 cells (*E. coli*, ATCC 47076) were cultured, collected and disrupted. They were subsequently subjected to reduction, alkylation, acetone precipitation, trypsin digestion and strong cation exchange (SCX) fractionation. The detailed protocol was described in Chapter 3. All the peptide fractions were then desalted and analyzed by RPLC-QTOF to collect MS/MS spectra. In total, 88,306 spectra were collected and searched with both Mascot (version 2.2.1) and X!Tandem (The Global Proteome Machine Organization, 2007.07.01, version Cyclone) using the same search parameters including: enzyme, trypsin; fixed modifications, carbamidomethylation (C); variable modifications, acetylation (N-term), ammonia-loss (N-term C), pyro-Glu (N-term Q), pyro-Glu (N-term E), and oxidation (M); precursor mass error, 30 ppm, fragment mass error, 0.2 Da, maximum missed cleavages, 2.

6.2.3 Human Data Set

Similar to the *E. coli* data set, SU-DH-L1 cells[14] (A human lymphoma cell line, ATCC) were cultured, harvested, disrupted by cell lysis buffer and subjected to acetone precipitation, reduction, alkylation, protein reverse-phase fractionation and trypsin digestion. Then RPLC-QTOF-MS/MS analysis was performed on all the desalted peptide fractions to collect MS/MS spectra. Overall this human data set contained 46,494 MS/MS spectra and was searched by Mascot and X!Tandem with the same parameters aforementioned in the *E. coli* data set section.

6.2.4 Validated *E. coli* Data Set

This data set consists of 34,993 experimentally validated spectral identifications. Each of them was examined using a ^{15}N-metabolic labeling

validation process. This approach is described in detail in Chapter 3. Briefly, unlabeled and ^{15}N-labeled *E. coli* spectra were collected and further compared by overlaying the spectra of unlabeled and labeled matches of the same peptide sequence for validation. Two cut-off filters, one based on the number of common fragment ions and another one on the similarity of intensity patterns among the common ions, were developed and applied to the overlaid spectral pairs to reject incorrectly assigned spectra. The search parameters used were the same as those for the *E. coli* data set.

6.2.5 Databases

Target-decoy search strategy proposed by Elias and Gygi in 2007[8] was applied by searching the MS/MS spectra against two separate databases (target and decoy databases) to calculate the global false discovery rate (FDR). The target databases used for *E. coli* and human data sets are *E. coli* K12 proteome sequences (size ≈ 2 MB, 4,339 sequences) and international protein index human database (IPI human database, version 3.68, size ≈ 48 MB, 87,061 sequences), respectively. The construction of a decoy database in this study was to simply reverse all the protein sequences found in target database.

6.2.6 Percolator Processing

The raw search results from Mascot (*.dat file) and X!Tandem (*.xml file) were imported to Percolator (version 2.01). The original result files were then parsed. Scoring features were extracted accordingly and sent to Percolator for further training steps.

In Mascot Percolator, the following features were used: precursor mass, charge, score difference between the best and second best match, precursor mass error, fraction of variable modification sites that was modified, the number of missed cleavages, fragment mass error, total intensity of the spectrum, total intensity of peaks that were used to identify a peptide, relative total intensity of peaks that were used to identify a peptide, fraction of ions that were matched in an ion series.

X!Tandem has a different scoring scheme than Mascot. Instead of reporting the probability of a PSM being random, X!Tandem first plots a distribution of calculated hyper scores from a specific search and then extrapolates E-values to provide an statistical evaluation for each identification. It was therefore found that the features extracted from X!Tandem were not exactly the same as the ones from Mascot or SEQUEST. All the features used in the X!Tandem Percolator were listed and explained in Table 6.1. As shown in Table 6.1, all those features can be categorized into 3 different groups, spectral quality, scoring and PSM statistics. In the category of spectral quality, all three features represent the intrinsic quality of an MS/MS spectrum regardless of its peptide assignment. In the category of scoring, all of the 8 features come from the original X!Tandem scoring algorithm and are used to measure the reliability of the sequence to spectrum assignment. Lastly, all the features in the PSM statistics category involve the information that are not directly used by X!Tandem but still might indicate the difference between true and false PSMs. They all can be switched on or off based on different requirements of applications.

6.2.7 Comparison

In order to evaluate the performance of X!Tandem Percolator, various comparisons among Mascot, Mascot Percolator, original X!Tandem and X!Tandem Percolator were carried out. For the experimental validated data set (see section 6.2.4), factual FDR was proposed to measure the error rate of search results. Because of experimental validation of sequence assignments, correct and incorrect PSMs can be physically isolated by comparing the X!Tandem or Percolator result with the validated result. Therefore, factual false discovery rates (FDRs) were accurately calculated by dividing the number of total PSMs with the number of incorrect PSMs. For real shot-gun proteomic data (see section 6.2.2 and 6.2.3), q-values were used instead. By definition, q-value is the minimal global FDR at which a PSM is accepted. Using widely accepted programs as references (e.g., Mascot and Mascot Percolator), receiver operating characteristic curves (ROC curves) were plotted to examine the number of PSMs at different factual FDR or q-value levels to demonstrate the improved performance of X!Tandem Percolator.

6.2.8 X!Tandem Percolator

X!Tandem Percolator was developed based on a mix of PHP and Perl scripts in an Apache server and it is integrated directly into X!Tandem. A simple click at the interface was all that needed to begin calculating features, re-ranking peptide identifications, assigning statistical values and exporting results in Percolator. X!Tandem Percolator can calculate features at 57 PSM per second on a quadcore 3.20GHz Phenom II 955 AMD processor.

Table 6.1 Complete List of Features Extracted from X!Tandem Search Results.

Index	Feature name	Feature type	Feature description
1	Mass	Spectral quality	The observed mass $[M+H]^+$
2 - 5	Charge	Spectral quality	Four Boolean features indicating the charge state
6	MaxI	Spectral quality	The maximum fragment ion intensity
7	PSMSumI	Spectral quality	The $\log_{10}$ value of the sum of all of the fragment ion intensities
8	Log(E)	Scoring	The $\log_{10}$ value of the expectation value for the peptide identification
9 - 10	IonScore	Scoring	The summed intensities of different types of fragment ions (y, b ions)
11 - 12	IonNo	Scoring	The number of peaks that matched between the theoretical and the test mass spectrum
13	HyperScore	Scoring	X!Tandem's score for the peptide Identification
14	NextScore	Scoring	The HyperScore of the second best peptide match of the spectrum
15	DeltaScore	Scoring	The difference of HyperScore between the best and the second best peptide matches
16	DeltaM	PSM statistics	The difference in calculated and observed mass (Th)
17	RelDeltaM	PSM statistics	The relative difference in calculated and observed mass (ppm)
18 - 19	IonFrac	PSM statistics	The fraction of fragment ions being matched in an ion series (y, b ion series)
20	MissClea	PSM statistics	The number of missed internal enzymatic (tryptic) sites
21	FragError	PSM statistics	The average mass error of all the fragment ions
22	AnnoPeaks	PSM statistics	The fraction of high intensity peaks being annotated as fragment ions
23	ModNo	PSM statistics	The number of variable modifications
24	ModFrac	PSM statistics	The fraction of modifiable residues being found modified (variable)
25	EnzN	PSM statistics	Boolean value: is the peptide preceded by an enzymatic (tryptic) site?
26	EnzC	PSM statistics	Boolean value: does the peptide have a C-terminal enzymatic (tryptic) site?
27	PepLeng	PSM statistics	The length of the peptide identification

6.3 Results and Discussion

6.3.1 Feature Selection

Even though Percolator is a semi-supervised learning method that does not need to construct a manually curated training set, in order to train a support vector machine, a variety of specific features that are capable of discriminating between true and false PSMs is still required. Understandably, the choice of features is vitally important. First, a list of features was composed that are supposedly capable of differentiating true and false PSMs. As shown in Table 6.1, there are 27 features that are able to measure the intrinsic quality of spectra and the quality of PSMs. The rationale is that spectral level information might indicate what types of MS/MS spectra (e.g., precursor charge states and fragment ion intensities) are more likely to lead to correct identifications. The features in the scoring category, such as Log(E) values, measure how reliable PSMs are individually from the perspective of the search engine. Finally, the features in the category of PSM statistics suggest how likely one PSM is to lead to a true identification from a global perspective. Different from the original and Mascot Percolator, it was decided not to include features that exploit protein-level information because protein-level feedback might change the score distribution of spurious PSMs, resulting in PSM misclassification[15, 16].

In order to determine if these features will actually discriminate between true and false positives, a set of experimentally validated PSMs from a previous study on *E. coli* cell lysates were used as standards (see Chapter 3). The

advantage of using this data set was that every one of the 34,993 PSMs was both correctly annotated and experimentally validated. The stringent experimental validation procedure that all those PSMs had undergone made them a set of highly reliable peptide identifications and consequently a perfect positive training set to extract features from. In addition, unlike the limited number of peptides from several standard proteins, the greater number of peptide identifications from this *E. coli* cell lysates made the analysis more statistically robust. In terms of the negative training set, the matches from the decoy search would be good representatives of false PSMs. By comparing the listed features from both results, the difference between true and false identifications were visualized in box plots. This comparison was important for all the chosen features, considering that they are the key to differentiate PSMs.

For all the features in the scoring category, features 8 – 15 in Table 6.1, in principle, should be able to show good discriminatory power. For instance, feature 8, log(E), is the $\log_{10}$ value of the expectation value for the peptide identification calculated by X!Tandem. As the main X!Tandem score to determine the reliability of PSMs, it was not surprising to see that true PSMs have a distinctively different distribution of log(E) than decoy PSMs. As shown in Figure 6.1A, the notches around the median log(E) values on the box plots for both true and decoy PSMs did not overlap with one another, indicating that the median log(E) value in true PSMs was significantly lower than that in decoy PSMs. Similarly, as indicated in Figure 6.1B, D, E, F and H, the same conclusion can be drawn for features including hyperscore, delta hyperscore, y ion score and y ion number. On

the other hand, as Figure 6.1C and G suggest, the difference between true and decoy PSMs was not as great as the rest of the features in Figure 6.1. However, this does not mean that they should not be included in the X!Tandem or Percolator algorithm. On the contrary, they can both be explained. In terms of the feature called next highest hyper score, considering that the next best match is supposed to be a random match, it was reasonable to see little difference between true and decoy PSMs for this feature. As for the b ion number, the little difference between true and decoy PSMs can be ascribed to the intrinsic fragmentation preference of tryptic peptides. Because most tryptic peptides have a more basic C-terminus, during the fragmentation process, y ions are preferably observed. Consequently, the number of observed b ions is much less than that of y ions in true PSMs and very similar to the number of randomly matched ions (y or b ions) in decoy PSMs.

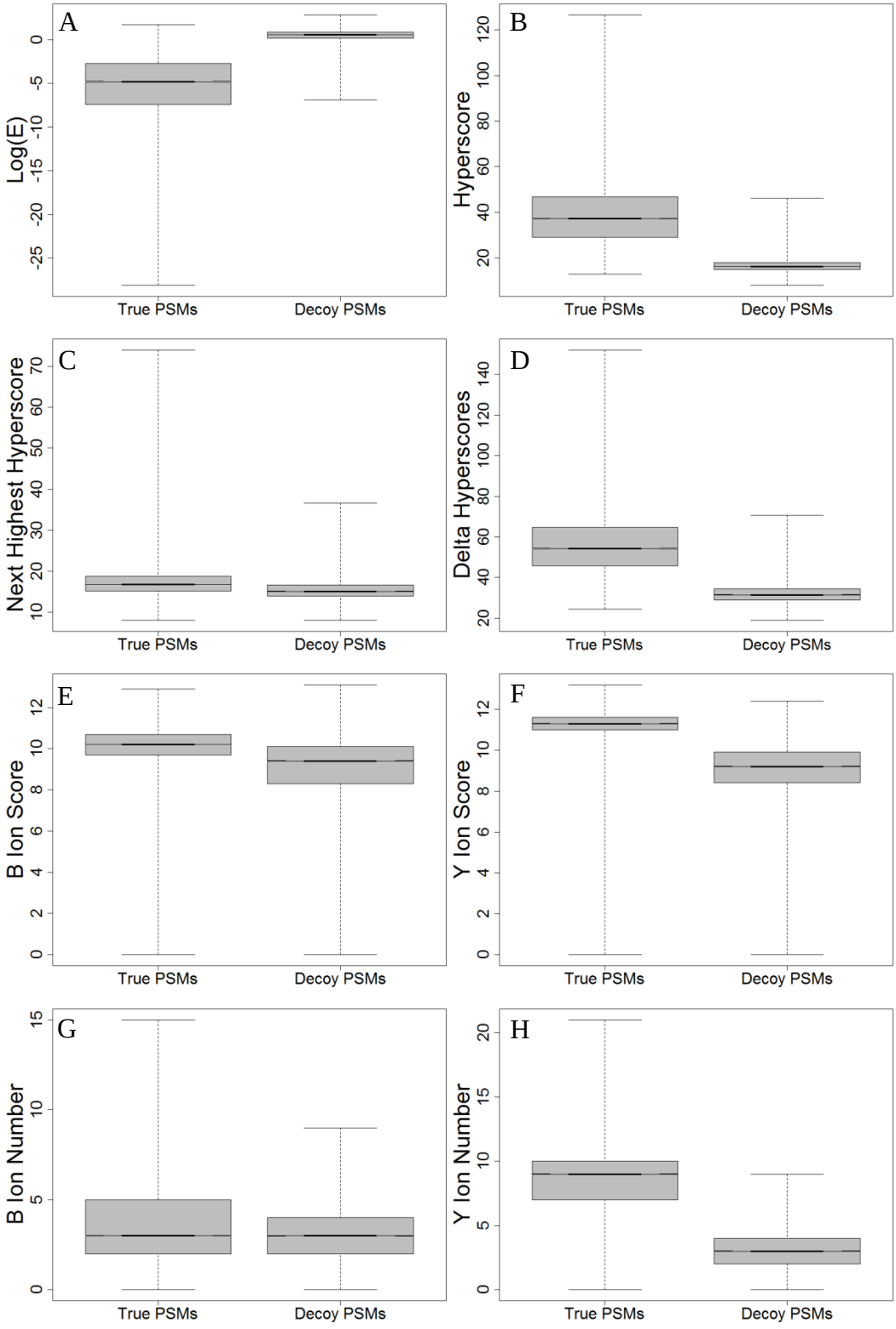

Figure 6.1 The difference between true and decoy PSMs in scoring features.

Apart from the features that were directly used by the X!Tandem algorithm, a set of features that indicate PSM statistics were also used. Take the feature fragment error as an example. In its definition, it represents the average absolute mass error of all the matched fragment ions. Generally, in a sequence database search, in an attempt to fully annotate an MS/MS spectrum, a wide mass tolerance window for fragment ions is cast. Unfortunately, it inevitably increases the possibility of random matches. However, in principle, the true fragment ions should have a smaller average absolute mass error than false ones, especially when MS/MS spectra were acquired with decent accuracy (e.g., QTOF data). Based on this concept, the fragment error should be able to provide a unique perspective to pin down false PSMs, which are primarily identified by detection of false fragment ions in MS/MS spectra. As illustrated in Figure 6.2F, the notches around the median values on the box plots did not overlap with one another, indicating that the median value of the fragment error in true PSMs was significantly lower than that in decoy PSMs. In addition, a new feature called AnnoPeaks was created, which was defined as the fraction of high intensity peaks (at least 70% intensity of the most intense peak) that were matched as fragment ions. Since peptides fragment in a reasonably predicable manner, most high intensity peaks in an MS/MS spectrum should be accounted for in a true PSM. Meanwhile in a false PSM fragment ions are matched by random peaks regardless of their intensities. Thus the feature AnnoPeaks should be able to provide another distinct perspective to distinguish true and false PSMs. By visualizing the distribution of AnnoPeaks values for both true and decoy PSMs in the box plot

(Figure 6.2B), the non-overlapping notches around their median values suggested that AnnoPeaks was indeed a good feature to discern true and false PSMs. Similarly, as indicated in Figure 6.2C, D, E, G and H, the same conclusion can be drawn for features including Precursor Mass Error, B, Y IonFrac, MissClea, ModNo and ModFrac (see their definitions in Table 6.1).

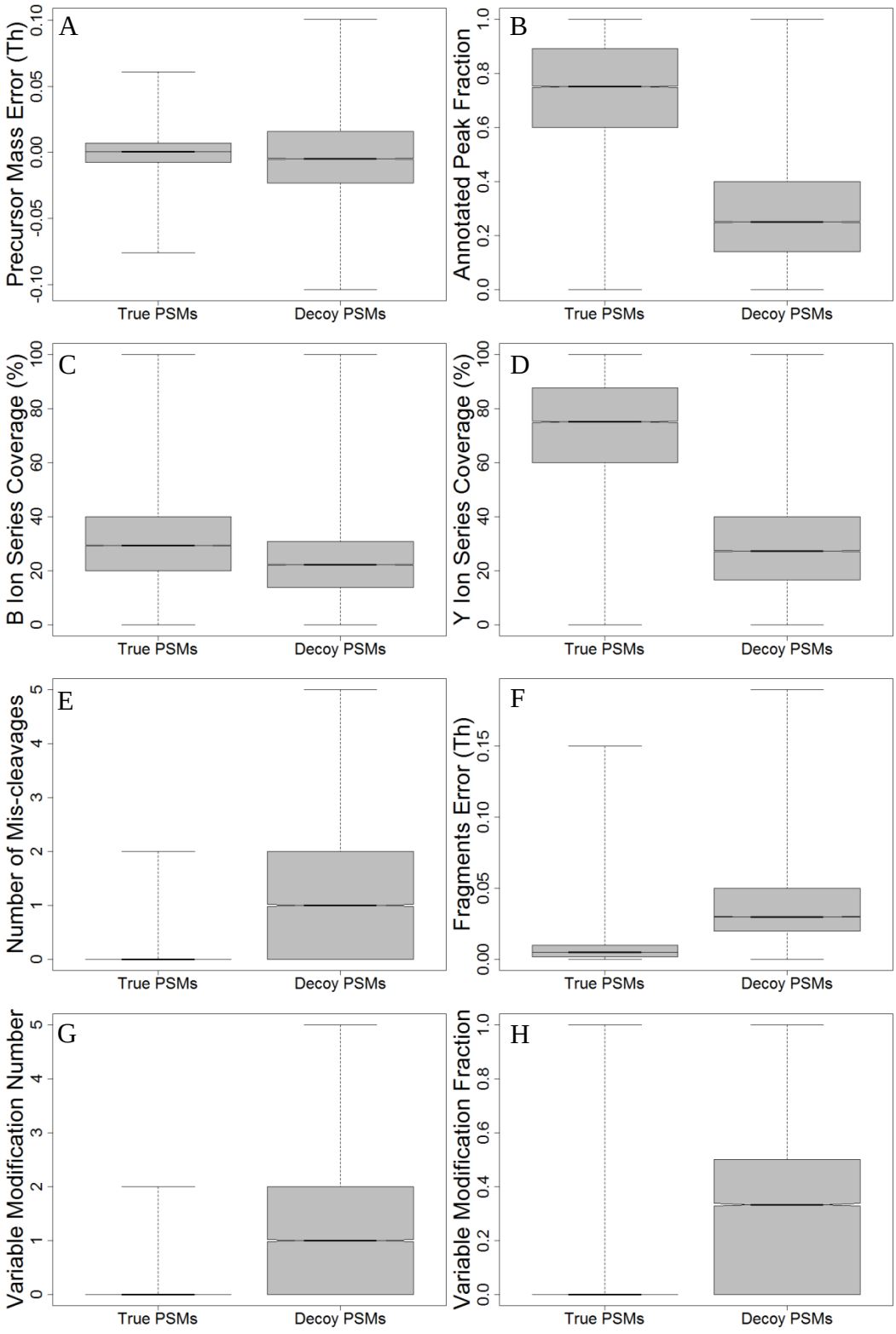

Figure 6.2 The difference between true and decoy PSMs in PSM statistics features.

In the category of spectral features, it was conventional for Percolator programs (SEQUEST and Mascot Percolator platforms) to include them in the process of differentiation. The rationale was that spectral level information might indicate what types of MS/MS spectra (e.g., precursor charge states and fragment ion intensities) are more likely to lead to correct identifications. Since those features are not a direct measurement of the quality of a sequence assignment, they might not be powerful discriminators to differentiate true and false positives when used individually. That was exactly what has been observed in this study. (see Figure 6.3). As shown in Figure 6.3A, the distributions of the quasi-molecular ion masses were very similar between the true and decoy PSMs. The median values of their quasi-molecular ion masses were not distinguishable. Similarly, based on Figure 6.3B and C, the same conclusion can be drawn for both the total fragment intensity and the maximum fragment ion intensity features. However, even though the spectral features individually were not very indicative in terms of differentiating PSMs, they still might contribute to the task when working collaboratively with each other or with features from the PSM category. The best way to examine their contribution is through feature removal analysis on real shotgun proteomic data. The *E. coli* data set (see section 6.2.2), was searched with X!Tandem and run on Percolator, eliminating one subset of features at a time. As shown in Figure 6.4, spectral features did make a contribution in the process of differentiation, even though individually they did not show strong discriminatory power. The number of estimated correct PSMs at a q-value of 0.01 was summarized, as well as the percentage decrease in estimated correct PSMs

relative to using all the features. As shown in Table 6.2, removing spectral quality features led to a 1% drop in performance, while removing PSM statistics features resulted in a 9% drop. However, comparing to the original X!Tandem result, X!Tandem Percolator equipped with all the features can significantly improve the number of PSMs.

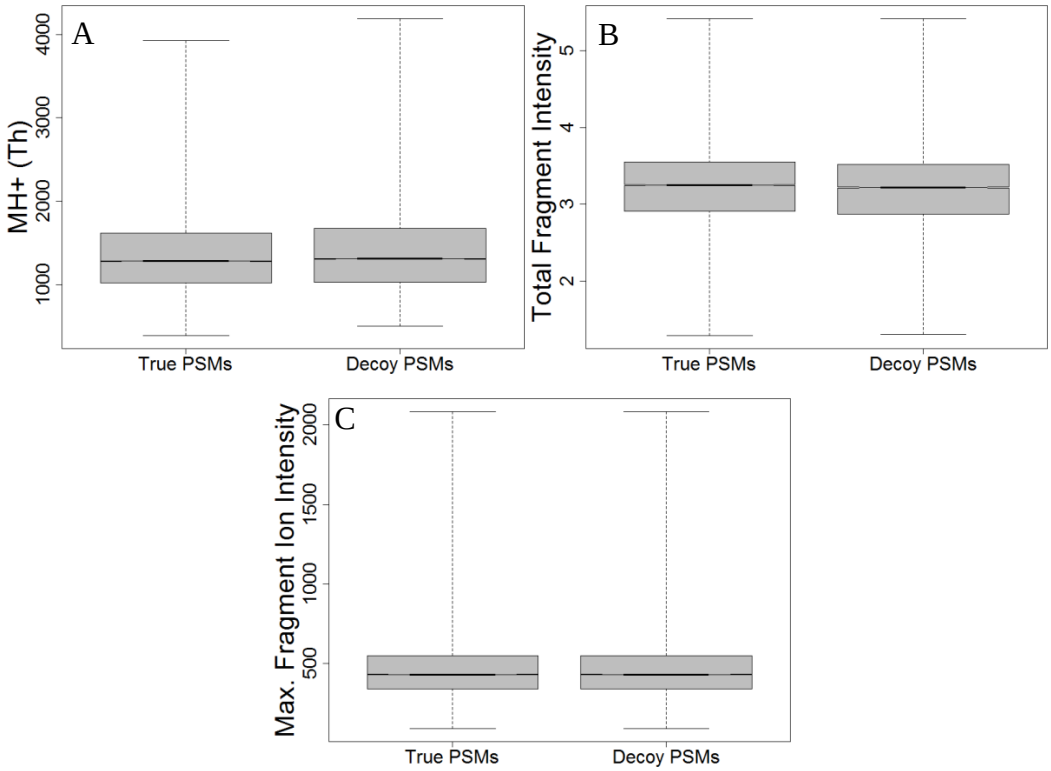

Figure 6.3 The difference between true and decoy PSMs in spectral features.

Table 6.2 Performance of X!Tandem Percolator When Fed with Different Features.

	Number of estimated correct PSMs	Drop in performance
All features	11478	-
Spectral quality features removed	11314	1%
PSM statistics features removed	10442	9%
Original X!Tandem	8786	23%

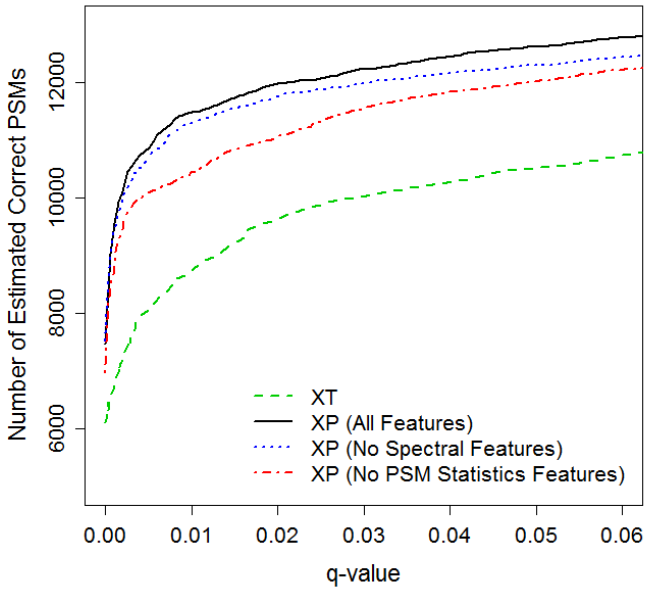

Figure 6.4 Performance of X!Tandem (XT) and X!Tandem Percolator (XP) when fed with different features.

6.3.2 Performance on Validated Data Set

After building a list of useful features based on an experimentally validated data set, a comparison was carried out on the performance of X!Tandem and X!Tandem on the same data set to examine X!Tandem Percolator's accuracy. Unlike most normal shot-gun proteomic data in which only a part of spectra (30 to 70%) are identifiable, all the un-identifiable spectra were filtered out in this validated data set. This means that any robust statistical tool should be able to recover close to 100% of all the pre-validated PSMs. In fact, it was found that X!Tandem alone was able to re-identify 98.9% of all the pre-validated PSMs when a lenient E-value threshold (E-value = 1) was applied. Since X!Tandem Percolator was designed to minimize the number of false positives and negatives of the original X!Tandem result, it was reasonable to observe a similarly high recovery rate with better sensitivity and specificity trade-off. As expected, for

X!Tandem Percolator the recovery rate was found to be 99.9% when a q-value of 0.4% was applied. Due to the advantage of validated data set, factual false discovery rates (FDRs) were calculated by dividing the number of PSMs with the number of incorrect PSMs (see section 6.2.7). When comparing the X!Tandem result with X!Tandem Percolator results at different factual FDR levels, the superior sensitivity and specificity that X!Tandem Percolator provided (see Figure 6.5)was apparent. In fact, at the factual FDR level of 1% X!Tandem Percolator estimated the q-value to be 1% as well, indicating accurate statistical assessment of X!Tandem Percolator.

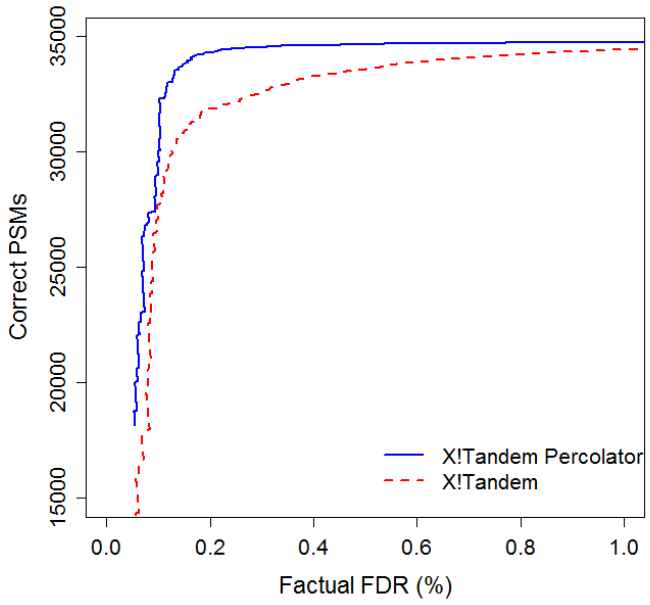

Figure 6.5 Performance comparison between X!Tandem and X!Tandem Percolator at different factual FDR levels.

6.3.3 Example Experimental Data

In order to examine X!Tandem Percolator's robustness, Percolator was applied to X!Tandem search results for large *E. coli*, and human lymphoma cell

data to demonstrate the superior sensitivity and specificity. Using our validated data set, true and false positives were detected and the factual FDR was calculated accordingly. But in typical shot-gun proteomic data sets, there is no validation of sequence assignments. Researchers therefore rely on the X!Tandem and Percolator programs to estimate the number of true and false PSMs and consequently calculate the q-value (the minimal global FDR at which a PSM is accepted) for each PSM.

6.3.3.1 Performance in Small Database

First, a performance comparison between Mascot, Mascot Percolator, X!Tandem and X!Tandem Percolator was carried out on the shotgun *E. coli* data set (see section 6.2.2). The *E. coli* system was selected because of its relatively simple proteome complexity (only about 4300 predicted proteins) and its popularity as a model system in proteomic studies. The data set was first searched by both X!Tandem and Mascot, and then processed by Percolator programs, respectively. Figure 6.6 shows the number of estimated correct PSMs for X!Tandem, Mascot, Mascot Percolator, and X!Tandem Percolator at different levels of q-values. As indicated in Figure 6.6, both Percolator programs offer much better sensitivity and specificity trade-off than Mascot and X!Tandem. To be exact, at q-value of 0.01, X!Tandem Percolator managed to identify 11594 PSMs, corresponding to 1391 proteins. Compared to the X!Tandem result (8875 PSMs and 1209 proteins), that was 31% and 15% increase in the number of PSMs and proteins, respectively. The same trend was also observed in the comparison between Mascot Percolator and Mascot, which had been reported by Brosh et al.[12]

as well. Overall, this result demonstrated the performance advantages of X!Tandem Percolator over the original X!Tandem scoring method when dealing with a simple proteomic system. Moreover, after the statistical analysis by Percolator, X!Tandem and Mascot results became more much agreeable to one another. At q-value of 0.01, the percentage of PSMs appeared in both Mascot and X!Tandem results constituted only 46% of total PSMs that were identified by both search engines. However, the percentage increased to 82% after Percolator was applied. This can be attributed to the better sensitivity and unified statistical assessment of Percolator.

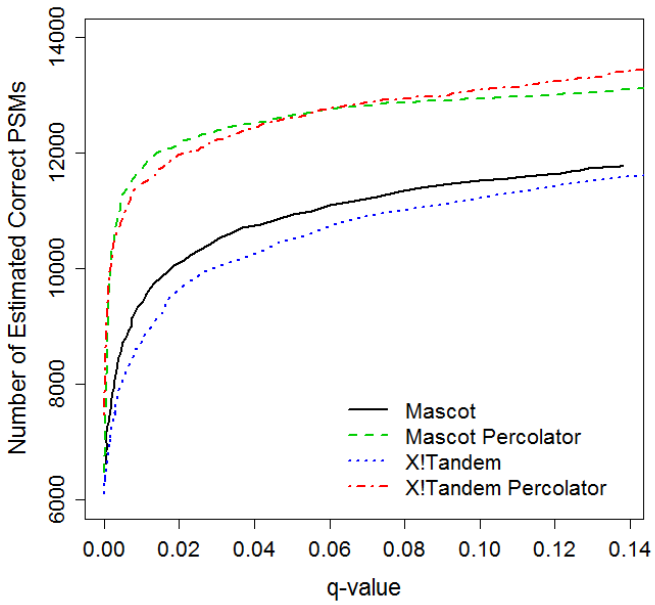

Figure 6.6 Performance comparison between Mascot, Mascot Percolator, X!Tandem and X!Tandem Percolator on the shotgun *E. coli* data set.

6.3.3.2 Performance in Large Database

The same analysis was applied to the human data set (see section 6.2.3) to see how X!Tandem Percolator would respond to the searches with a much larger

proteome database (87061 protein sequences). Understandably, a much larger proteome database provides more combinations of amino acids. It is therefore an even bigger challenge to differentiate true and false PSMs. Figure 6.7 shows the number of estimated correct PSMs for X!Tandem, Mascot, Mascot Percolator, and X!Tandem Percolator at different levels of q-values for the human data set. As indicated in Figure 6.6, both Percolator programs still offer much better sensitivity and specificity trade-off than Mascot and X!Tandem. In fact, at q-value of 0.01, X!Tandem Percolator was able to improve the number of PSMs and protein identifications of the original X!Tandem result by 52% and 29%, respectively. The improvement on the human data set was even greater than the improvement on the E. coli data set, indicating that Percolator was less easily influenced by the complexity of proteome databases.

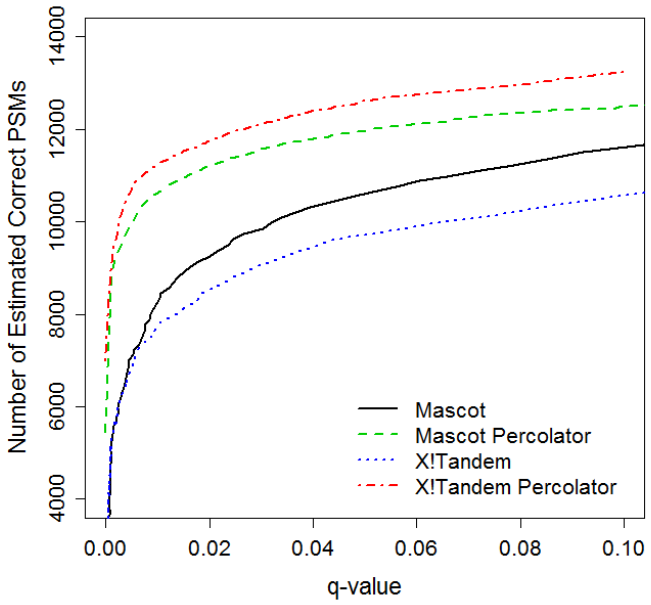

Figure 6.7 Performance comparison between Mascot, Mascot Percolator, X!Tandem and X!Tandem Percolator on the shotgun human data set.

6.3.3.3 Sensitivity to Search Space Change

Sometimes, relaxed searching parameters are chosen by researchers in order to match as many peptide sequences as possible in a proteomic study. For instance, a wide precursor mass tolerance window is often cast so as to capture all the potential identifications. Contrary to the original intention, it is often noted that when relaxed searching parameters were set up for a sequence database search engine, a noticeable drop in the number of PSMs is often observed. This is simply due to the fact that increased search space creates more possible random matches. In order to avoid a decrease in accuracy, sensitivity is often sacrificed. In this study, in order to test how well X!Tandem Percolator is able to handle this issue, different precursor mass tolerance settings, including 15, 30 and 500 ppm was used in X!Tandem. After searching the human data set with those settings, results were processed by Percolator. As shown in Figure 6.8, as the precursor mass tolerance setting increases from 15 ppm to 30 ppm, an increase in PSM number for both X!Tandem and X!Tandem Percolator is obvious. In fact, at q-value of 0.01, the improvement in PSM and protein identification were 8% and 1% for the X!Tandem result, and 11% and 4% for X!Tandem Percolator result, respectively. It indicated that 30 ppm was more appropriate mass accuracy window for this data set. When increasing the setting from 30 ppm to 500 ppm, little change is observed for either X!Tandem results or X!Tandem Percolator results. However, at q-value of 0.05, another commonly used threshold, a decrease in X!Tandem performance (5% less PSMs, 2% less proteins) can be easily spotted in Figure 6.8. At the same time, almost no decrease in X!Tandem

Percolator is observed. It is suggested that X!Tandem Percolator is a highly robust statistical tool and less easily influenced by search space increase.

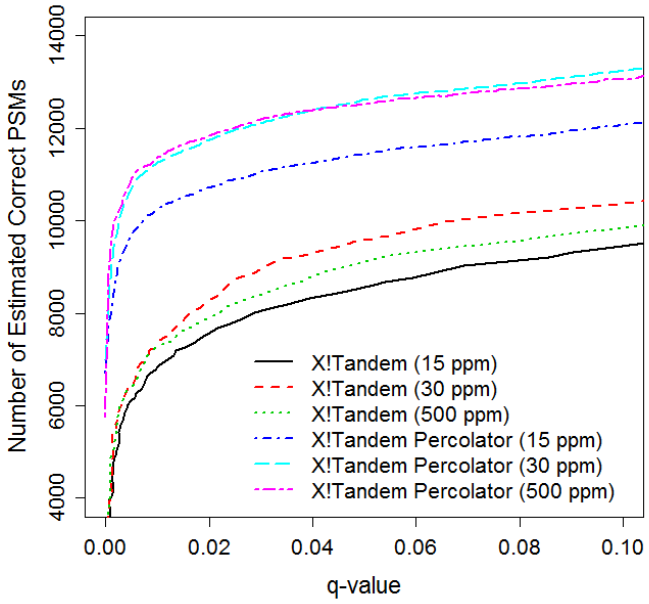

Figure 6.8 The influence of precursor mass tolerance setting on X!Tandem and X!Tandem Percolator.

6.4 Conclusions

Percolator was previously shown as a very robust classifier that can dramatically improve sensitivity on various search engines, such as SEQUEST[10] and Mascot[12]. In this study, an interface has been built for Percolator and X!Tandem, a very popular open-source search engine.

To successfully integrate Percolator with X!Tandem, a large number of features that define the quality of PSMs were first created. Since an experimentally validated data set provided the opportunity of isolating the true PSMs from search results, by comparing the features from true and decoy PSMs,

the individual discriminatory power of each feature was carefully examined. Moreover, a feature removal analysis was also performed to demonstrate the collective contribution of different subsets of features.

X!Tandem Percolator was applied to shotgun proteomic data, including the *E. coli* and human data sets. Under various conditions, including different sizes of databases and relaxed search parameters, X!Tandem Percolator always seemed to substantially outperform the original X!Tandem, showing a similar or even better performance of Mascot Percolator. Overall, it demonstrated that better classification of true and false PSMs can be achieved when multiple factors are working collaboratively instead of just using one scoring metric.

6.5 Literature Cited

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198-207.

(2) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. *Nat. Methods* **2010**, *7*, 681-685.

(3) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III *Anal. Chem.* **1997**, *69*, 767-776.

(4) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(5) Fenyoe, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.

(6) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(7) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods* **2007**, *4*, 787-797.

(8) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(10) Kaell, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

(11) Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.

(12) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.

(13) Xu, M.; Li, L. *J. Proteome Res.* **2011**, *10*, 3632-3641.

(14) Wu, F.; Wang, P.; Zhang, J.; Young, L. C.; Lai, R.; Li, L. *Mol. Cell. Proteomics* **2010**, *9*, 1616-1632.

(15) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. *J. Proteome Res.* **2009**, *8*, 3737-3745.

(16) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111-1120.

Chapter 7

Conclusions and Future Work

In MS-based proteomic studies, the correlation between the spectra and peptide identifications is a crucial step as well as a challenging task. Even with the triumph of various search strategies and algorithms, there is still room for improvement. The overall goal of this thesis research is to develop a spectral searching strategy capable of identifying peptide sequences from MS/MS spectra with high sensitivity and accuracy.

In Chapter 1, key terms are explained and sample preparation, instrumentation, isotopic labeling methods, search algorithms and statistical analyzing tools are discussed as related to mass spectrometry based protein sequencing.

In Chapter 2, a shotgun proteome analysis method was developed and successfully applied to the identification of proteins from thousands of cancer cells. Since a small number of cells were used, cells were disrupted using a detergent (NP-40) containing solution instead of French press to minimize sample loss. The lysed cells were subjected to acetone precipitation, followed by cautious washing with cold acetone and solubilization in NH_4HCO_3. After trypsin digestion, the resultant peptide mixture was analyzed by RPLC-ESI MS/MS. To achieve the best mass spectrometric performance, the gradient slope of RPLC profile was optimized according to the sample amount injected into the column.

It was shown that this method could identify an average (n=3) of 167 ± 21, 237 ± 30, 491 ± 63, and 619 ± 59 proteins from 500, 1000, 2500, and 5000 MCF-7 cells, respectively. To demonstrate the potential use of this method for generating proteome profiles from cancer cells isolated from human blood, MCF-7 cells were spiked into a healthy human blood sample and this mixture was processed and then subjected to antibody tagging of the MCF-7 cells. The tagged cells were sorted and collected using flow cytometry. The proteome profiles of small numbers of cells isolated in this way were found to be similar to those of the MCF-7 cells. This work illustrated that proteome profiling of a small number of cells isolated from blood can be achieved. By comparing the obtained profile to a standard profile, cell typing might also be possible, which may prove to be useful for cancer diagnosis or prognosis.

In this work, Mascot, a sequence database search engine, was used to identify peptide sequences from MS/MS spectra. After manually validating numerous peptide-sequence matches, it was clear to me that the Mascot identity threshold strategy was unduly conservative to match all the identifiable peptides in a data set. An example was shown in Figure 7.1. Even though the entire series of y ions were identified, the peptide was still considered insignificant by Mascot as the Mascot ion score was lower than the Mascot identity threshold. This issue of sequence database search strategy definitely leads to reduced sensitivity of peptide and protein identification. In order to further improve the proteome profiling of a small number of cells, a more sensitive strategy for correlating peptide and MS/MS spectra should be used.

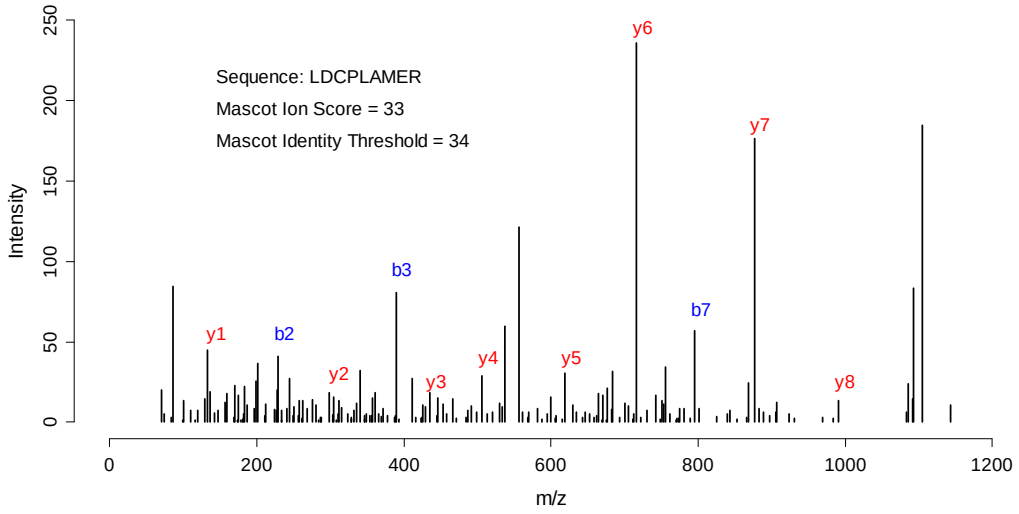

Figure 7.1 An example of a highly reliable peptide match that is considered insignificant by Mascot.

In Chapter 3, an attempt was made to develop spectral search strategy for peptide and protein identification using MS/MS spectra. It was demonstrated that spectral searching strategy is more sensitive than sequence database searching strategy. One should keep in mind, though, that an accurate identification by spectral searching strategy is built on the premise of a reliable MS/MS spectral library. In Chapter 3, ^{15}N-metabolic labeling approach was first developed to experimentally validate all the peptide matches from Mascot search results. In this experimental validation approach, the MS/MS spectra of the unlabeled peptides and their ^{15}N-labeled counterparts can be overlaid and their fragmentation patterns and mass shifts due to nitrogen number differences can be readily compared to validate the spectrum-to-sequence matches. For spectral validation, two cut-off filters were developed. One was based on the number of common fragment ions found in the overlaid spectra. The second filter was based on the similarity of the

fragmentation patterns of the unlabeled and labeled peptide pairs. A similarity score was calculated by using the fragment ion intensity dot-product.

Using *E. coli* K12 proteome analysis as an example, a highly confident MS/MS spectral library was constructed by using those experimentally validated peptide matches. This library consists of 9,302 unique spectra (unique sequence and charge state) from 7,763 unique peptide sequences. Finally, a spectral searching algorithm called SpecMatching was developed to utilize this spectral library. In analyzing a digest of an *E. coli* extract using both Mascot and SpecMatching, it was shown that SpecMatching provided better sensitivity and specificity even with this small-size spectral library.

However, it does not mean that sequence database searching strategy should be completely replaced by spectral searching strategy. In fact, spectral searching and sequence database searching should be considered complementary to one another. Sequence database searching is more suited for discovery-oriented studies of which the objective is to identify novel peptides or modifications. In contrast, spectral searching is a more effective way to detect previously identified peptides. Given the complementary nature of those two strategies, in the future they can be combined in tandem as a better peptide identification method. For instance, sequence database searching can first be applied to a reference sample to construct a reference spectral library. Then the more sensitive spectral searching strategy can be applied to subsequent experiments to detect and even quantify the same set of interesting proteins under different experimental conditions. This

tandem method is particularly suited for studies involving numerous samples and replicates, such as clinical studies.

In Chapter 4, ^{18}O-labeling method was chosen as an alternative to validate peptide matches from a human cell digests as it is too costly to apply metabolic labeling strategy on human samples. In combination of the precursor inclusion strategy, all the identified PSMs (peptide-sequence matches) had a chance to be validated by their ^{18}O-labeled counterparts. It successfully resolved the issue that some highly confident unlabeled PSMs cannot find a labeled counterpart. As a result, a large number of unlabeled PSMs from human cell lysates have been validated and are ready to be compiled into a spectral library for future usage. With the application of the precursor inclusion strategy, categorizing all the PSMs into correct and incorrect PSMs became possible with the application of three quality filters. Compared to Chapter 3, where only correct PSMs can be isolated, this categorization has an advantage. It made the calculated false discovery rate (1 – validation rate) no longer an estimation but an accurate representation of the error content in a result.

With those validated PSMs, the performance of commonly used search engines (Mascot and X!Tandem) and two popular statistical approaches (PeptideProphet and Percolator) were carefully examined. In Chapter 4, it was found that PSMs identified by multiple tools had lower error rates than the ones identified by only one tool. Then by comparing the numbers of validated PSMs at the same validation rate, it was found that it was better (more validated PSMs) to

embrace the overlapped PSMs from multiple statistical tools than simply to raise the score threshold. Next, it was confirmed that the unreasonably rigorous identity threshold was the cause of the poor sensitivity of Mascot when searching large space. Apart from using Mascot Percolator or Mascot PeptideProphet, using global FDR estimated by target-decoy strategy as a threshold instead of using Mascot identity threshold was a possible way to improve its sensitivity. Besides, it was also confirmed that X!Tandem (with refinement function on) is not compliant with the target-decoy strategy. Moreover, applying Percolator or PeptideProphet to original search results could truly improve the number of true PSMs while maintaining a relatively low error rate. Finally, the investigation on the performance of all five statistical tools revealed that Mascot Percolator outperformed the other four statistical tools.

In Chapter 5, experimentally validated PSMs from Chapter 4 were further used to examine the validity of identifications from Mascot and X!Tandem results on the protein level. Thanks to the advantages of experimental validation, it became possible to isolate true and false positives on both peptide and protein levels. With the numbers of true and false positives, one can readily calculate (not estimate) the true global FDRs of peptide and protein identifications. It was demonstrated that a low global FDR on the peptide level cannot guarantee a low global FDR on the protein level. If the goal of one's study is protein identification (e.g., biomarker discovery and proteome profiling), a simplistic global FDR control on the peptide level is insufficient to gauge the reliability of protein identifications. In this study, it was found that the commonly used "two-peptide

rule" can in fact significantly improve the reliability of protein identifications but is unduly conservative.

In order to recapture the correct single-hits eliminated by the "two-peptide rule", a further categorization of all the single-hits discovered two subgroups: homologous single-hits (HSHs) and strict single-hits (SSHs). It was observed that HSHs were as reliable as multi-hits and thus recommended to be treated as such. With respect to the majority of single-hits, SSHs, two straightforward solutions were proposed to discern the true positives from false ones. If one has access to two search engines (e.g., Mascot and X!Tandem), compare SSHs from one result with all the protein identifications from the other. The SSHs that can be found in both results are highly reliable and should be deemed as confident protein identifications. If there is only one search engine available, a two-stage threshold approach seems to be a rational choice. In the first stage, a relatively lenient score cut-off (e.g., significance threshold of 0.05 in Mascot or maximum E-value of 0.05 in X!Tandem) is chosen for the all the PSMs. Next, use all the peptide matches that pass the score cut-off to infer protein identifications. While keeping all the multi-hits and HSHs intact, apply a more stringent score cut-off (e.g., significance threshold of 0.01 in Mascot and maximum E-value of 0.005 in X!Tandem) on all the SSHs to improve their collective reliability. Using either approach, more protein identifications can be identified than the overly conservative "two-peptide rule" without sacrificing the global protein FDR.

In Chapter 6, the validated PSMs from Chapter 3 was used to build an interface between Percolator, one of the most powerful statistical evaluation tools and X!Tandem, a popular open source sequence search engine. The key to this successful interfacing was to generate a large number of features that define the quality of PSMs. Since experimentally validated data set provided the opportunity of isolating the true PSMs from search results, by comparing the features from true and decoy PSMs, the individual discriminatory power of each feature was carefully examined. Moreover, a feature removal analysis was also performed to demonstrate the collective contribution of different subsets of features. Then X!Tandem Percolator was applied to shotgun proteomic data under various conditions, including samples from different species, different sizes of databases and relaxed search parameters. X!Tandem Percolator always seemed to substantially outperform the original X!Tandem, showing a similar or even better performance of Mascot Percolator.

Even though Chapter 5 and 6 are not directly related to develop the spectral searching algorithm or construct spectral libraries, yet those two projects are extensions of experimental validation approaches, the main quality control on the reliability of spectral libraries. In fact, from Chapter 5, a better understanding on how to deal with single-hit proteins was achieved. It will definitely be useful in the near future when inferring protein identifications from peptide matches by spectral searching strategy. From Chapter 6, a significant sensitivity boost was obtained for X!Tandem, a sequence search engine. It was not possible without the correct PSMs from our ^{15}N-metabolic labeling validation experiment in Chapter 3.

In turn, the enhanced sensitivity of X!Tandem will also benefit the construction of spectral libraries by creating more peptide-sequence matches from raw MS/MS spectra.

Finally, it is obvious that the success and applicability of the spectral searching strategy depends heavily on both the reliability of PSMs and the proteome coverage of the spectral libraries. We envisage the use of experimental validation strategies to construct MS/MS spectral libraries of various organisms for proteome analysis with improved sensitivity and specificity. To generate a comprehensive MS/MS spectral library of a model organism, such as *E. coli*, more detailed proteome analysis, such as the use of cellular fractionation (e.g., membrane-bounded vs. plasma) and protein separation (e.g., based on molecular weights), followed by multi-dimensional LC-MS/MS, will be needed. In the near future, more protein profiling and experimental validation work will be done to achieve the goal of building up comprehensive and reliable spectral libraries.

Meanwhile, in the future, a fully automated procedure based on the protocol will be created to make library construction easier. Besides, the spectral searching algorithm, SpecMatching, needs a user-friendly interface. Ideally, a web server containing spectral libraries of model organisms and a spectral searching tool will be constructed to accommodate the need for spectral searching based research. Without a doubt, as the advance of mass spectrometry technology and the rapid accumulation of MS/MS data, the spectral searching strategy will also improve and gain more useful features, such as faster speed and surrounding

informatics support. Given their complementary nature, in the future applications, spectral searching will be bundled with sequence database searching strategy as a follow-up analysis to increase the number of peptide and protein identifications. Besides, the concept of spectral searching can be easily applied to monitor the performance of the LC-MS/MS instrumentation to ensure day-to-day and lab-to-lab consistency and quality.

Furthermore, compared to unprocessed spectral identifications, spectral libraries with consensus spectra will serve as superior gold standards for bioinformatic studies, such as in understanding of peptide fragmentation mechanism, because the library spectral identifications are consolidated from many replicates, properly de-noised and deemed highly reliable.

Lastly, the peptide identifications that are compiled in spectral libraries have been proven to be ionizable, detectable and identifiable. In combination of the valuable knowledge of the fragmentation and MS/MS spectra, we can easily choose proper candidates and develop selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) assays for biomarker verification studies. It is reasonable to expect that spectral libraries will play an important role in the developing platform of targeted proteomic studies and connect discovery-based and verification-oriented proteomics.